统计如何说谎

[1]

自学书院　译着

2014



[1] http://www.surgerycenterrecruiter.com/wp-content/uploads/2012/08/Leukemia-Statistics-lg.jpg

译言

看了第一本有关统计谎言的著作 *How to Lie with Statistics by Darrell Huff, 1954*，立论精辟，虽然书中一些例子已经过时，理据依然对照现在的「统计误世」年代。计算机软件又引进了一些新工具和误区。考虑之下，为保留原作面貌，选择译本每章分为两部份。第一部份翻译原书（略有删节，省掉没有历史背景资料很难明白的例子），第二部份选译补充材料，主要参考如何利用统计数据撒谎（WikiHow）、统计学〈维基百科〉、统计误用〈维基百科〉、Misleading graph〈维基百科〉以及其他网页数据。

Creative Commons 有限版权制度面世已经十年，全球有一百三十多个国家和地区已有本土化的 Creative Commons 条款。Creative Commons 条款适用于任何创作成果：大如维基百科，YouTube 视频、Flickrs 相片集，小如个人网志，都可以是以 Creative Commons 条款发表的学习和应用材料。如各路英雄一呼百应，本着知识共享的精神，壮大 Creative Commons 的范畴，互相支持，互补互助，网上的知识源泉定必波澜壮阔。

华文世界的 Creative Commons 发展，有是有，但比诸其他语言，实在落后于人。「革命尚未成功，同志还需努力。」

关于「统计学」的 Creative Commons 著作，我只找到刘彦方和陈强立的《思方网 ：统计与图表》，如高人有其他发现，请告知。

自学书院
2014 年 6 月

统计的重要



复杂的现代社会离不开调查和统计。相关人员收集、整理、归纳、分析数据和发表结果，广泛应用在自然科学、社会科学和人文科学，也用于决定工商业及政府政策。日常生活躲也躲不了的广告也每每以统计数据引导消费者。

统计是为面对不定状况制定决策提供方法的科学。统计学和机率论的关系异常密切，事实上任何统计问题的研究都必涉及机率论的运用，后者实为前者的主要工具。统计可以是利用现有数据或通过调查取得数据。除非**母体群**[2](population)规模特小，调查可以覆盖全部，一般调查是以取样方式进行：搜集小量数据（样本 sample）的数据以估计、预测和研究母体群。

统计陷阱带来的负面影响可大可小。基于错误统计的政策可能差之毫厘，谬以千里；医学的统计陷阱可能要数十年后才被纠正，招致人命损失。近代广告特多统计数字引导误导消费者。

要了解统计的诸多陷阱，先看看一般统计的流程。

利用现有数据的统计主要是案头作业，这方面的陷阱亦见诸调查统计。要搜寻未知的数据，抽样调查是最常用的搜集方法。

一般而言，统计作业的步骤如下：

1. 决定调查主题。
2. 决定收集数据的方法：(a)书面作业或(b)调查：面对面访问，邮寄问卷、电话访问或混合运用。
3. 界定(a)书面作业的范围或(b)抽样调查的母体群。
4. 决定(b)抽样使用的母体群清册：如电话号码簿、会员名单、户籍资料等。
5. 决定(b)抽样方式：随机抽样、分层抽样、系统抽样或分段抽样。

---

[2] 亦作 parent population, universe；有译为「总体、母体、母群 」。

6. 决定(b)样本大小；若需分层，需决定分层方式及各层样本大小。
7. (b)进行抽样，选取样本元素。
8. 设计(b)收集数据的形式；设计调查问卷，预试。
9. (a)汇集资料；(b)执行调查，向样本收集反馈。
10. (a)和(b)数据检误、处理及分析。
11. (a)和(b)发表结果。

从上可见，每一步骤都涉及人为因素和诸多可操控手段。无论是什么形式的统计，都可能出错；这可能是意外，也可能是故意，构成统计陷阱。

有三种谎言：谎言，该死的谎言和统计数字。___Benjamin Disraeli

总有一天，有教养的公民能读能写，也要有统计思维。___H. G. Wells

我们不知道的那些事情不会让我们陷入困境，
而是我们知道但并非如此的事情。___Artemus Ward

数字与统计

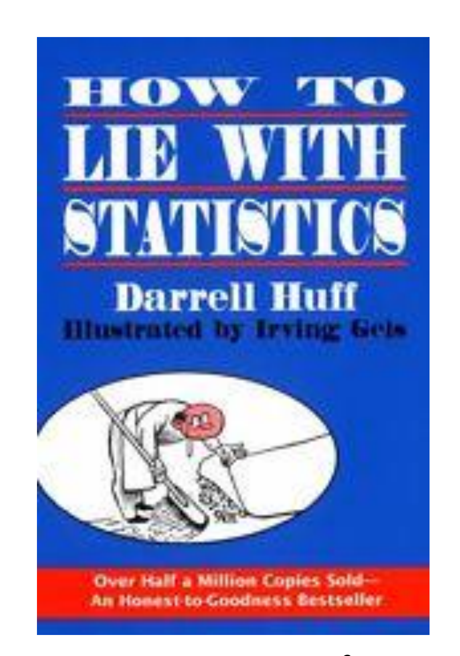
「多数人对于数字具有先天的畏惧感，是有演化的根源；因为人类存活在地球有几十万年，大多数时候是几十人、最多百来人的小族群过着狩猎采集的生活，全部家当两只手就可拿着走，因此不需要用上什么数字，对成千上万的大数字更是没有概念。只有在近一万年来，人类实行农业生活后，人类社会的规模与财富不断累积成长，才开始出现对数字的需求，也才有天赋异禀之士发展出各式各样的数学。

虽然多数人对数字可能无感，但冰冷的数字还是要比感性的言语可靠。统计是整理大数字的科学方法，如果是因为不懂统计，或吃过统计的亏，就把统计与谎言并列，可说是因噎废食，也算另一种人的偏见吧。」

___〈潘震泽：人类天生的缺陷：数字盲〉

引文说「把统计与谎言并列」是「另一种人的偏见」。相信没有人会把全部统计看作为谎言，但统计有误区，也不能否认有人利用统计说谎。统计有什么误区？如何说谎？这是本书的主题。

统计如何说谎？



Darrell Huff, 1954[3]

目录

[3] 原文：How to Lie with Statistics by Darrell Huff, 1954。译本略有删节，减掉一些不懂历史背景很难明白的过时例子。

序言

神圣古老的英国度量衡制度快要取消，英寸和英尺的时代快要结束；盖洛普民意以一贯方式测试人们对取而代之的公制的认识，发现大学程度的男女有 33 %从未听过公制。

然后一份周刊的读者调查宣布读者有 98 %知道公制。对此，报刊吹嘘它的读者群比一般人「更懂行」 。

两项民调如何能够有这么明显的差异？

盖洛普调查员精心挑选了公众的样本并约见会谈。这家报刊儿戏和经济地依靠读者填写和邮寄问券。

由此不难猜测大部分不知道公制的读者根本没有兴趣填报和邮寄问券，自动不参加调查。用统计术语来说，这样的自我选择只会产生具偏见或不具代表性的样本，多年来导致许多许多误导性结论。

多年前的冬季，十多位独立调查员报告抗组织胺药片的数量，各人都发现药片治愈大多数感冒病例。

于是广告和医疗产品的热潮开始炒得火热。这是基于人们对灵丹妙药的永恒希望，也没有超越统计数据去看看长久以来我们已经知道的事实。幽默作家 Henry G. Felsen 不是医学权威，很久之前已指出适当的治疗可以在七天治愈感冒：只要多休息，置诸不理，一星期就会好转。

因此，你读到的和听到的平均值、关系、趋势和图表并不是表面的真实无误，背后可能有更多或更少的讯息。

在追求事实的文化中，统计的秘密语言是如此吸引人，实则是用来炒作，夸大，混淆和简单化。在报告社会和经济趋势、企业经营状况、「民意」调查和人口普查的大量数据，统计方法和统计术语是必要的，但报告者用辞必须诚实和易于了解，读者也知道用辞的意思，才不会陷于语义的无稽之谈。

科普文章滥用统计数字，几乎排挤了在半明不亮实验室日以继夜辛勤研究的白袍英雄。统计资料粉饰许多重要的事实，犹如扑粉化妆，上油涂漆。精心包装的统

计胜于希特勒的「大谎言」；只是误导，但难以追究。

这本书是如何使用统计数据来欺骗的读本，可能看起来太像骗子手册。也许我的好理由是模拟退休窃贼出版的回忆录等同如何挑锁和消声毁迹的研究生课程：骗子都知道这些技巧，老实人必须为了自卫而学习。

第一章　有内置偏差的样本

桶内有红豆白豆，有一种办法肯定各有多少：倒出来点数。

有一个更简单的办法算出有多少红白豆。假设桶内的红豆白豆是相同比例，拿出一把豆子，只计数这一把。就大多数目的而言，如样本足够大和选择正确，这足以代表整体。但如两方面有偏差，其准确度可能远远及不上聪明的猜测，只不过是所谓科学精确的虚言。样本因为选择的方法有失偏颇，或过小，或两者兼而有之，会导致谎言，也就是我们读到或以为我们知道那些很多结论背后的可悲事实。

样本如何出现偏差？请看一个极端的例子。假设要发问卷调查，其中包括以下的问题：「你是否喜欢回答问卷调查？」之后收回的问卷极有可能得出这样的结论：「典型的样本人口绝大多数喜欢回答问卷调查」，其准确度可计算至几个小数点。这是什么一回事？当然是因为回收的问卷已排除了大多数可能回答「不喜欢」的问卷，调查问卷已掉在废纸篓。即使原始样本中十有八九是「不喜欢」那帮人，这些「错误」已排除在外。

现实生活中是否有这样的有偏样本？当然有。

不久前，报刊和新闻杂志报导在过去十年有约四百万美国耶教旧教信徒改信新教。消息来源是跨宗派《耶教导报 *Christian Herald*》编辑 Daniel A. Poling 牧师的调查。《时代》周刊总结这故事：

> 《导报》的数字来自美国新教牧师，2,219 位牧师填报（发出问卷 25,000 份），呈报共有 51,361 前旧教教徒在过去十年加入新教。Poling 依样本估算在十年有 4,144,366 名旧教教徒改信新教。Will Oursler 主教写道：「即使估算有出入，全国数字不会少于二，三百万，极有可能接近五百万。」

虽然《时代》有报导调查中超过 90 %牧师没有填报问卷，但错过了指出这事实的重要性，依然精神可嘉。要彻底摧毁这调查，唯一要注意的合理可能性是大多数牧师扔掉问卷是因为没有改信教徒的数字可以呈报。

利用这假设和 Poling 采用的相同数字（181,000 名牧师），可以另行推算。他的调查涵盖 181,000 牧师的 25,000 人，呈报 51,361 人改信新教；如调查涵盖全部 181,000 牧师会得出有约 370,000 人改信新教。

这样的粗糙方法得出非常可疑的数字，但至少是一如前一数字值得信任；那个全国数字是修正数字的十一倍，因此更引人注目。至于 Oursler 主教对误差的自信，如果他发现了一种方法来纠正未知大小的误差，将会造福统计界。

在这背景下，多年前有另一则新闻报导，当时的币值较高：耶鲁大学学生平均年收入有$25,111。很棒！

且慢。这令人印象深刻的数字是什么意思？这是否表明如果子女进读耶鲁或牛津，剑桥，你和他不用年老时上班？

第一眼看过去，这数字有两个疑点：令人惊讶的精确，也不大可能这样的令人称羡。

只有极小可能性可以精确得知任何散漫群体以往任何时候的平均收入，更不要说精确至$111。除非收入全来自薪金，很少人能如此精确知道自己的年收入。有这样收入的人往往会分散投资。

此外，这个可爱的平均数无疑是源于耶鲁毕业生的自报收入。即使耶鲁大学在 1924 年校风纯朴，但不能保证四分之一世纪后这些毕业生都如实自报收入。被问及他们的收入，有些人因虚荣心或乐观夸大了。其他人少报，尤其是担心纳税申报，不想在任何其他文件留下自相矛盾的数据。谁知道税务局会否看到？吹嘘和低估这两种倾向可能相互抵消，但其实是不可能的。其一倾向可能远远强于另一，但不知道是哪一个。

先说一下：常识告诉我们这数字几乎不是真相。这信息表示一些人的「平均收入」是$25,111，而这些人的实际平均收入可能较接近一半。现在看看信息可能来源的最大误差。

常识告诉我们不可能在二十五年后与当年的全部毕业生保持联络。有人已往生，有人地址不详。

那些有通讯地址的，很多人不会回答问卷，特别关乎相当个人的资料。对于某些类型的邮件问卷，5-10%的反应已是相当高的。这一个调查的回报率应该比这更好，但肯定不是 100%。

因此，这收入数字源自有已知地址而又乐意填报个人收入的毕业生。这是否具代表性的样本？也就是说，是否可以假设这群组的收入是相等于没有参加调查（没有地址或不愿回报）的另一群毕业生？

在耶鲁名录，那些毕业生「地址不详」？是否那些赚大钱的华尔街巨子，公司董事，制造业及公用事业主管？不，富人的通讯地址不难查得到。即使他们忽略了联系校友办公室，从名人录和其他参考刊物找出他们的通讯地址应是轻而易举。二十五年后失联的毕业生，按常理猜测应是那些毕业后事业不顺的毕业生：文员，技工，流浪汉，失业酗酒汉，仅堪糊口的作家和艺术家。可能几个人的收入总和才可攀上$ 25,111 的收入水平。他们不那么经常参加旧生联谊活动，可能有些人甚至不能负担旅费。

谁会把问卷撵到垃圾桶？不能肯定，但公平的猜测至少是很多人没有挣多多的钱可以自我吹嘘。这有点像新员工发现第一份工资单夹着纸条，建议他保密工资数额，不与同事交换机密数据。这家伙会告诉老板：「别担心，我和你一样为此感到羞耻。」

看来很清楚样本省略了最有可能压低平均水平的两组。那个$25,111 数字开始为自己解释。这只适用于有己知地址，又愿意公开本人收入的特殊群体。这还要假设他们是说真话的君子。

不要轻易作出这样的假设。抽样调查的一个品种即是所谓「市场调研」，其经验表明根本不能作出假设。有一项市场调查的关键问题是：你家看什么杂志？结果列表和分析显示很多人喜爱高端的 *Harper's*，这虽然不算是曲高和寡，但至少算得是中上阶层口味；并没有很多人自认是低俗杂志 *True Story* 的读者。然而，出版商的数字很清楚表明 *True Story* 的发行量有几百万份，而 *Harper's* 只有几十万。调查的设计人员自我解困：也许我们问错了对象。但事实不是这样。调查在全国各地街上访问。那么唯一合理的结论是很多受访者回答这些问题时没有说实话。调查只是发现了人们在装腔作势，装模作样。

最终发现，如果想知道某些人看什么杂志，查询是没用的。更好的办法是从他们家里买入旧杂志，这中自有信息。

只需数算《耶鲁评论》和《爱情周刊》的册数。即使这样也不能确实知道人们在看什么，只是知道他们接触什么。

同样，读到有报导一般人（最近听的很多，大部份不可信）刷牙每天一到两次（我随意取一个数字），这有什么问题？谁能知道这些事情？　女生看了无数广告，印象中以为不刷牙是社会罪行，她会否向陌生人承认她不经常刷牙？这样的统计只意味着人们对刷牙的说法，但没有弄清楚人们刷牙的频率。

谚语有云：河水向下流，不高于源头。嗯，这似乎是可能的，如果有泵站帮忙。同样真实的是抽样调查的结果不会优于样本本身。数据经通过层层统计处理，过滤为小数点平均值，调查结果开始蒙上可信的光环，但仔细看看采样就可以否定这假像。

可信的采样报告必须采用具代表性的样本，即是已去除每一偏见的源头。上文的耶鲁数字顿见毫无价值。许多报刊和杂志报导犯下同样错误，没有什么意义。

有一次，精神科医生报告谓几乎每个人都是神经质。这样的说法除了破坏「神经质」一词的任何意义，倒不如看看这位医生的样本，也就是说这位精神科医生一直在观察什么人？原来，他是从观察他的病人得出这启发性结论；这个「样本」根本不能作为总体人口的样本。正常人不会看心理医生的。

阅读不要囫囵吞枣，可以避免学习了一大堆表里不一的东西。

值得铭记无论是有形或无形来源的偏差都会破坏样本的可靠性。也就是说，即使不能找到可证实偏见的来源，只要有偏差的可能性，对结果也应保持一定程度的怀疑。

一项例证是 1936 年《文学文摘》月刊的著名惨败。月刊的一千万名电话用户和月刊订户调查曾准确预测 1932 年的总统大选。1936 年，月刊汇集同一名单的反馈，编辑部放心预测罗斯福只有 161 选举人票，对手 Landon 得票 370。这样本名单久经测试，怎会有偏差？当然有偏差；无数高校论文和其他事后研究发现：在 1936 年有财力安装电话和订阅杂志的人不是全体选民的横截面。这个富裕组群是特殊的组群；这是一个有偏差的样本，因为大多数样本是共和党选民。这样本选择 Landon，但全体选民却不以为然。

基本样本被称为**随机**(random)，在母体群中被选中纯粹是偶然；统计人员指全体为「母体群」，样本是其中部份：索引卡每十个名字选一个，每批纸张取五十张，在闹市每二十名行人采访一位。(但请记住，这不是这个国家或城市人口的样本，只是当时闹市区域的样本。一项民意调查的访问员声称可在火车站「找到各种人等。」必须指出她的误区：例如，带着小童的母亲可能比例不足。)

随机样本的测试是这样的：是否每一个名字或事物在整体中有平等机会成为样本？

纯随机抽样[4]，是唯一可以利用统计理论检查而又令人有全面信心的统计方法，

[4] purely random sample

但其多种用途的成本昂贵和执行困难，令人望而却步。民意调查和市场研究这些普遍领域几乎都采用更经济的替代品：**分层随机抽样**[5]。

要得出分层抽样，先把母总群按已知**盛行率**[6]比例分为**组群**[7]。麻烦从此开始：所知的比例讯息可能不正确。调查员按指示访问多少名黑人（以收入阶层细分百分比），多少名农民等等；这些组群必须均分为四十周岁之上和之下。

听起来有层有次，但实际情况是怎样？大部分时间调查员不会弄错受访对象是黑人或白人。收入方面会多犯错。如何界定农民：在农场兼职又在城市上班应如何分类？即使年龄也可能带来一些问题，避重就轻的办法是只选择明显低于或超过四十周岁的受访者。在这种情况下，样本有偏差，没有包括三十多岁和四十多岁的年龄组。你不能全赢。

考虑以上各点，应如何在分层内得出随机样本？最明显的先找出全体人口的姓名列表，从中随机选择；但成本太昂贵。所以访问员走到街上（偏误是忽略了留在家中的人们），或是在白天挨家挨户访问（偏误是忽略了上班族），或换到晚上访问（忽略了电影迷和夜游人）。

意见调查的操作，归结到底是对有偏见来源的持久战，所有著名的民调机构时时刻刻都在作战。阅读调查报告时，必须记住这是必然败北的战斗，从来没有赢过。「英国人有 67%反对…」或其他类似的结果，先要问问这 67%是什么英国人。

美国著名的人类性学研究者金赛博士[8]与他人合着的《金赛报告[9]》：《男性性行为》（1948 年）及《女性性行为》（1953 年）。《报告》无疑是划时代的研究，但样本远远不是随机，令人不安。样本名单有极大偏差：女性受访者 75%有大专以上学历，男性受访者有颇大比例是囚犯(25%)或男妓(5%)[10]。更严重的误区是样本大幅度倾向有性暴露狂的受访者；乐意向访问员诉说性历史的人，其经历大大有异于对访问员说不的沉默寡言群体。

布鲁克林学院 A. H. Maslow 在金赛之前有一项研究，参与的女学生许多后来也志愿参与金赛的研究；Maslow 发现这些女生普遍是较为性成熟和独立特行。这证实了人们对金赛研究的质疑。

阅读《金赛报告》或任何有关性行为的较近期研究时，要懂得适可而止：即是不

---

[5] stratified random sampling
[6] prevalence
[7] group
[8] 金赛博士 Alfred Charles Kinsey，1894-1956
[9] Kinsey Reports
[10] 译文略有补充，参考维基百科。

要过份阅读。任何基于采样的研究都突显这样的误区，尤其是大型调查的主要报告浓缩为摘要形式更可能变得如此。

首先，像《金赛报告》这样的研究至少涉及三个层次的抽样。上文已指出母体群（第一层次）的样本并不是随机，因此可能不特别代表任何母体群。同样重要的是要记住任何问卷可能只是许多可能问题的其中一个样本（第二层次）。受访者的答案只不过是响应那问题的个人态度和经验的样本（第三层次）。

类似金赛的性研究和其他调查都发现访问员的身份会影响调查结果。在二战期间，美国全国民意研究中心派出两位员工访问南方城市的五百名黑人。一位调查员是白人，另一位是黑人。

访问员提出三个问题。其一是「如果日本征服美国，黑人会得到更好或更坏待遇？」黑人访问员回报受访者有 9%回答「更好」。白人访问员得到同样的响应只有 2%。黑人访问员回报受访者有 25%回答「更坏」。白人访问员得到同样的响应却有 45%。第二条问题以「纳粹德国」取代「日本」，结果也是类似。

第三条问题探讨可能是基于前两条问题显露的感情。「专心击败轴心国或致力让民主更好在美国发展；你认为那一项更重要？」黑人访问员回报 39%选答「专心击败轴心国」，而白人访问员回报 62%。

偏误是因为许多未知因素。最有效的因素可能是人们有给出令对方满意答案的倾向，因此阅读调查结果时要自我提醒。回答在战乱时对忠于国家的问题时，南方黑人会告知白人访问员动听的答案，而不是本人实际相信的答案，这是不足为奇。也有可能是不同访问员选择不同类型的对象接受访问。

在任何情况下，结果是很明显是一面倒偏误，毫无价值。各位可以自行判断有多少调查的结论是一样偏颇，毫无价值，而且没有测试揭示这些偏误。

如果你怀疑一般调查偏向于特定方向，一如《文学文摘》的错误，这可视之为相对证据：受访者比代表母体群平均组群偏向更有钱，受较多教育，有较多信息和较高警觉性，更美好的外观，更常规的行为以及较稳定的习惯。

很容易看到如何产生这此偏误。假设访问员被分派到某街角完成面试。眼前两位仁兄似乎都适合要求的类别：第一位是四十处的城市黑人，不修篇幅；另一位穿着干净工作服，体面整洁。为了尽快完成访问任务，访问员更有可能向后者打招呼。全国各地的访问员都做出类似的决定。

自由派或左翼圈子对民调最反感，普遍认为民调一般被操控。这种观点的背后事实是民调结果往往不符合那些思想不保守人士的意见和愿望。他们指出民意调查似乎选上共和党，即使此后选民不是这样投票。

事实上，从上文所见，民调不是必然被操纵，刻意扭曲结果以制造假象。样本向这一致方向倾斜已是自动扭曲。

补充材料

选择母体群和抽样的误区

书面作业选用那些现有数据？调查选择那些母体群？全都影响统计数据。

即使母体群的界定符合「涵盖全体」的意思，如何从中抽样？[11]

- **简单随机抽样** simple random sampling，也叫纯随机抽样。从母体群 N 个单位中随机抽取 n 个单位作为样本，每一单位有相同机率被抽中为样本，即是每个样本单位被抽中的机率相等，每个样本单位完全独立，彼此没有一定的关联性和排斥性。简单随机抽样是其它各种抽样形式的基础，通常只是在母体群单位之间差异程度较小和数目较少时才采用。
- **系统抽样** systematic sampling，也称等距抽样。将母体群的所有单位按一定顺序排列，在规定范围内随机抽取一个单位作为初始单位，然后按事先规定规则确定其他样本单位。先从数字 1 到 k 之间随机抽取一个数字 r 作为初始单位，以后依次取 r+k、r+2k……等单位。这种方法操作简便，可提高估计的精度。
- **分层抽样** stratified sampling。将抽样单位按某种特征或规则划分为不同分层，然后从不同分层中独立、随机抽取样本。从而保证样本的结构与母体群结构比较相近，从而提高估计的精度。
- **整群抽样** cluster sampling。将母体群的若干个单位合并为组，形成抽样框，抽样时直接抽取，然后全部调查中选组群的所有单位。抽样时只需抽中抽样框，可简化工作量，缺点是估计的精度较差。

学术调查较多说明采用那种方法，但一般调查极少说明。以香港为例，有化妆品／牙膏等等广告标榜「90%（或高比例）女士／牙医选用…」；为适应法例要求，广告以极小白字标示数据来自什么什么调查。仔细一看，这些调查往往来自内部或母公司调查。这些数据应该是真实的，但这些「内部」调查是否随机？是否涵盖适当的母体群？牙医母体群是否包含全部注册牙医，或是参加广告方主办免费

[11] 这段落取自〈抽样〉《维基百科》，略有改写。

研讨会的参加者？「女士」是否局限于在该品牌化妆柜台浏览甚至购物的女士？

- 参考阅读：抽样与代表性

轻率概化和过度类化

统计的的特定总体不能代表母体群，即是轻率概化的谬误，例如调查只限于某政党党员和同路人而把结论概化为全民意见。

现实生活中的调查往往以电话进行，常有过度类化的谬误。如调查人员只致电手机（流动电话），而手机用户以年青人占大多数，这忽略了没有手机，只有家用电话的家庭主妇和老年人。这不是全民调查的正确取样。

抽样调查

常见的报导屡屡提到是次调查访问了多少人。大城市人口动辄千万，大国人口以亿计，究竟调查样本应有多少才有代表性？不懂统计学的人们少不免怀疑调查数千人是否取得数百万人的意见。完美公正的抽样和可信答案的调查，在数学上有误差范围，取决于调查的人数。

先要了解取样调查的两个重要术语：**置信区间**[12](confidence interval)和**置信水平**[13](confidence level)。置信区间也称为误差(margin of error)，即是调查报导时常提到的 ±X%。抽样误差本质上不是错误(mistake)，最完善的抽样统计程序和方法都无法避免抽样误差（除非刚巧每一个样本都具有和总体相同的特征，那另当别论）。

在既定的置信水平，影响其置信区间有三个因素：**样本大小**(sample size)、**百分比**(percentage)和**母体群规模**(population size)。

很明显较大的样本数量更能确保如实反映母体群的答案；也很明显最大范围的样本就是母体群全部，但这是不实际的，否则就无需抽样调查这回事。但在既定的置信水平，样本越大，置信区间越少；但这关系不是线性的，不是说倍增样本大小会导致误差率减半。

调查的准确度也取决于样本选取一个特定的答案的百分比。如样本 99%说「是」，1%说「否」，无论样本大小，错误的机会是微乎其微。然而，如答案的百分比是

---

[12] 亦有译为「信赖区间」。
[13] 亦有译为「信赖／信心水平／水平」。

51%对 49%，出错的可能性要大得多。

样本可能代表已知的国家或城市人口，或是不确切知道的准车主数目。机率数学证明如样本是母体群的百分之几，母体群的规模是无关紧要，除非母体群的规模偏小或是有既定特点的已知群体（例如某协会的成员）。

取样的黄金规律是「随机」，真正的「随机」。调查出错往往是因为取样不是随机。

以大家熟悉的盖洛普(Gallup)调查为例，看看「美国全国民意调查」是怎么抽样的。

> 无论是一次性或追踪性调查，盖洛普的取样是一千人，置信区间为±4%，置信水平为 95%。即使加大样本，误差不会有很大差异。
>
> 在收集数据后，盖洛普依据美国人口调查局的人口特征（性别、族裔、年龄、学历和地区）为每位受访者加权。
>
> 例如，调查一千名国民对总统的支持率为 50%，误差为±4%，即是支持率是在 46%至 54%之间。如样本扩大至二千人。误差可降至±2%，但成本倍增。
>
> 在决定样本多少时，调查机构必然要考虑成本。最准确的民意调查要涵盖全体国民，但这是不切实际。
>
> 「置信水平为 95％」的意思是如盖洛普进行一百次同样的调查，有九十五次的结果大致相同，只有五次不是在「46%至 54%」的范围。[14]

---

[14] http://www.gallup.com/poll/101872/how-does-gallup-polling-work.aspx
http://www.gallup.com/poll/File/125927/How%20Are%20Polls%20Conducted%20FINAL.pdf

Sample Size Calculator 是 Creative Research Systems 的网上公共服务，用来决定需要多少样本以反映目标母体群的精确结果。只要点选置信水平（95%或 99%），输入置信间距（误差）和母体群人数，就可以算出所需样本大小。[15]。

网页计算器要求输入以下的选择，如母体群的规模庞大或未知，可以留空。

决定样本大小 Determine Sample Size

置信水平 Confidence Level: ( )95% ( )99%

置信间距 Confidence Interval:

母体群 Population:

所需样本 Sample size needed:

计算置信区间 Find Confidence Interval

置信水平 Confidence Level: ( )95% ( )99%

样本规模 Sample Size:

母体群 Population:

百分比 Percentage:

置信区间 Confidence Interval:

不恰当的调查问题

问卷和电话调查都是由访问者提出问题，遣词用字能引导受访者给出有倾向性的答案。如二战期间的民意调查问题为：

- 德国已进占法国。美国应否参战？
- 日本已偷袭珍珠港。美国应否参战？

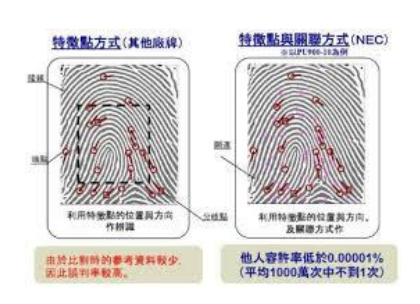
其中的预设立场显而易见。

另一陷阱是在诱导性提问加入导向「理想答案」的数据。例如：

- 中产家庭税务是多年新高，你是否支持扣减所得税？
- 国家提出庞大量赤字预算以应付迫切需求，你是否支持扣减所得税？

[15] http://www.surveysystem.com/sscalc.htm#one

法律与统计



一宗谋杀官司突显了严重的统计问答。虽然疑犯否认他在犯罪现场，但正面临控方提出的指纹证据。指纹专家在庭上被控方盘问：「被告人的指纹和其他人的指纹相同的机率是多少？」专家作答：「数十亿份之一。」辩方律师盘问：「在犯罪现场得到的指纹被错误识别为某人的机率是多少？」专家：「哦，大约是百份之一。」

指纹证据是事实，但识别指纹是判断，不是事实，是一门科学，并且由机率支配。[16]

〈视频〉Peter Donnelly: How juries are fooled by statistics 统计如何迷惑陪审团（中文字幕）。统计数字如何错判「杀婴案」。

---

[16] 数据源：http://www.nytimes.com/2013/05/26/opinion/sunday/when-numbers-mislead.html?_r=0

第二章　精心挑选的平均值

读者诸君不是势利小人，我当然不是地产代理。姑且假设你是势利暴富户，而我是地产代理。你打算在我熟悉的小区买房子。我打量一下，小心翼翼告诉你这小区的业主住客平均收入每年约一万英镑。也许这引起你的兴趣；无论如何，你决定买房子，也记住这年收入数目。势利暴富的你在告诉你的新地址时也不经意抛出这数字。

一年多后，我们又见面了。我是当区地方税缴纳人委员会的成员，要求小区的业主住客签署请愿书呼吁不要增加地方税或调低物业估值或公交票价减价，理由是这超出小区居民的负担，毕竟我们的平均收入每年只有£2000。

也许你会附和我和委员会的呼吁；你不仅势利，也懂得省钱。但你对年收入£2000的说法无法释怀：究竟我是现在或是去年说谎？

无论怎样，你不能怪责我。利用统计数据说谎就是这样的美好。这两个数字都是合法的**平均值** average，合情合法，都代表同样的数据，同样的居民，同样的收入。都是一样的。很明显，至少其中一个是误导，等同不折不扣的睁眼说瞎话。

我的诀窍是每次拿出不同类型的平均值；「平均值」有非常松散的定义。打算影响公众舆论或出售广告空间，这一招很管用，有时是无心之失，但往往是故意而为。要清楚明白「平均值」，先要知道是那种平均值：**平均数** mean，**中位数** median 或**众数** mode。

我抛出一万英镑数目时是想提出一个大数值：平均数是这小区所有家庭的收入的算术平均值：所有家户的收入总和除以家户数目。中位数是较小的数字：有一半家庭的收入多于£2000，有一半少于这数目。我也可以抛出众数，这是序列数据最常见到的。如这小区有最多家庭的年收入是£3000，每年£3000 就是众数。

在这种情况下，没有解释的「平均值」是毫无意义；收入数据一般也是这样。有另外因素乱上添乱：源自随着某些种类讯息的平均值差别不大，一般来说是无需着意区分。

如果有报告谓某原始部落的男性平均身高只有一米，你会对他们的体型有相当不错的见解，无需追问这是否平均数，中位数或众数，三者的数值都是差不多。（当然，如果你打算在非洲出售工作服，就要有比平均值更多的信息。这是关乎全距

range 和偏差 deviation，下一章详谈。 ）

处理诸如许多人性特点的数据时，不同的平均值是相当接近所谓「**正态分布**[17]」，以曲线表示其形状为钟型；平均数，中位数和众数都在同一点汇合。

因此，如描述人的高度，各种平均值是一样好；但如要描述某城市居民的收入，也许是由些微收入至二万英镑左右，某地可能有几个超级大户。超过 95%的居民的年收入是在五千英镑之下，曲线向左侧倾斜。这不再是对称的钟型，而是被扭曲，形状像小孩的滑梯，梯子急剧上升至一个高峰，滑下部分倾斜逐渐下降。平均数与中间数有相当距离。比对一年的「平均数」和「中位数」，其差异一目了然。

回到上文物业经纪就小区居民年收入抛出两个相差颇大的平均值，是因为分布明显倾斜。如居民大多数是小农户或打工一族或是年老退休人士，但有三位百万富翁周末业主，居民总收入的算术平均数是极大数值。几乎每个居民都在平均数之下。这是现实，但听起来像笑话或比喻而矣。

因此，读到企业或东主自白他员工的平均工资是什么什么，这数字可能有一些意思，也可能没有。如数字是中间数，意思是高于或低于中间数工资的员工各占一半。如果是平均数（如没有说明，一般是这个），所谓平均收入是£25,000 其实没有分开东主的得益和和低薪工人的工资。平均年薪£3,800 可能掩盖工人年薪£1,400 以及东主以高工资形式拿走大部份利润。

统计的语言伪术可以把坏事包装成为较好的外观。

三位合伙人开设一家小型制造企业。过去一年生意非常好，支付了九十名员工的工资（共£99,000）以及每名合伙人工资各£5,500 后，余下利润还有£ 21,000。如何描述这状况？为便于理解，可以利用平均值。

既然员工都做同样工作，薪酬没有太大差别，使用平均数或中位数都是差不多：员工平均工资　£1,100，合伙人平均工资和利润　£12,500

这看起来很可怕。换一种方式。三位合伙人分取利润£15,000（余下£6,000）。这一回以平均数计算员工和合伙人的工资：平均工资 £1,403，合伙人平均利润£2,000。

啊！这看起来更好：利润不足 6%。现在可以发布，张贴或在谈判中使用这些数

[17] normal distribution

据。

这相当粗糙的例子极度简化，但比对以会计之名做出的花招，这不算什么。在层次结构和复杂的公司，员工从打字员到年收几百万美元奖金的总裁，这样的手法可以掩盖各种各样的东西。

所以，看到平均工资的数字，首先要问：什么的平均？谁包括在内？美国钢铁公司曾表示其员工的平均周薪在不到十年上升了 107%。是的，他们没说错—— 但只要注意到十年前的数字包括众多兼职工人，这数字的意义就大打折扣。如某人去年是半职，今年是全职，他的收入增加一倍，但工资率其实是一样。

有报导美国家庭的平均收入是$ 6,940。要明白这个数字，先要知道何谓「家庭」以及是什么平均值。（以及谁这么说的？他怎么知道？数字是否准确？）

数字可能来自人口普查局。局方的报告全文说明这是中位数，「家庭」指「住在一起两个或两个以上有亲属关系的人」。报告还说明数据来自这样规模的样本，每二十个样本有十九个的估计是在±71 美元的范围。

这机率和误差率加起来是相当不错的估计。调查局人员有足够的技术和资源以相当精度程度完成取样研究。想必他们没有特别要遮掩的。但不是所有的数字都是是在这样的情况下快乐诞生，也不是伴随着任何讯息来说明如何精确或不精确。下一章详解。

看看《时代杂志》的〈发行人的话〉：新订户的年龄中位数为 34 岁，其平均家庭收入为每年$7,270。早前的调查发现旧订户的年龄中位数为 41 岁，平均收入为$9,535 美元。问题是为什么两次都给出年龄中位数，但刻意没有说明收入采用那种平均值。

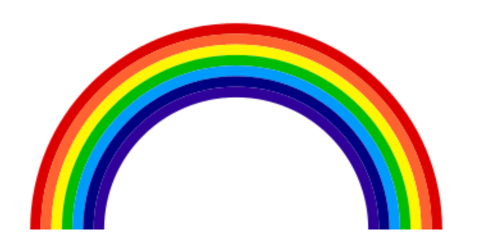
会否是用了平均值以表达较大数值，可以向广告商介绍读者群是如此富裕？

利用第一章的耶鲁旧生数据，猜猜是采用了那一种平均值。

补充材料

平均值的误区

讨论统计数据时少不免提到「平均值、平均数」。这名词的表面意思很明显：平均值就是大致居中的一个数值。但实际上有好几种平均值。



平均而言，彩虹是白色的。

※**算术平均值**(mathematical average/mean)是把所有数据加在一起，再除以总体的样本量计算。（3,3,5,4,7）这几个数值的算术平均值就是把总和(22)除以 5（因为有 5 个数值）；算术平均值是 4.4。

※**中位数**(median)是一组数值从低到高排列，恰好处在中间位置的那个数值。同上例子（3,3,5,4,7），中位数是 4，因为有两个数值(3,3)比它小，两个数值(5,7)比它大。

※**众数**(mode)是一组数值中最常见的数值。同上例子的众数是 3，因为出现了两次。

算术平均值看起来似是以上三种计算方式最简单的一种，但实际上不是这样。因为一组数据中如有过高或过低数值（极端的数值）对算术平均值产生很大的影响。

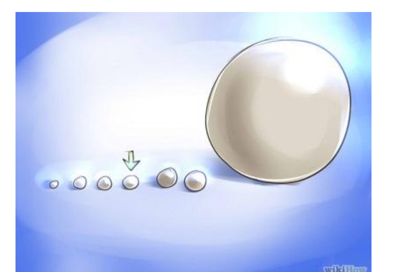
※例如，统计一个小区内 50 户家庭的收入。大多数家庭的收入是每年$40,000-60,000，但有一家每年收入是 5 百万元。如此这般的算术平均值因为 5 百万元这个数值而大大提高。

※如 9 个人各有 1000 元存款，第十个人只有 1 元存款，算术平均值是 900.10 美元。

比较可信的数据调查往往去掉最高和最低的数值才计算算术平均值。但不是每一

项调查都这么可信。除非看到所有数据或已去掉极值的说明，最好不要对这些数据照单全收。

中位数的误区



中位数容易有误区，因为和其他数据相比，这不是很明显过高或过低。中位数处于中间位置，很容易隐藏了那些很大或很小的数值。例如，数据是 0.1,1,2,3,4,5,3000，中位数是 3。

用中位数描述某事件随时间变化的程度时，容易遮掩事实。如过去九年每年涨价 3%，但今年涨价 20%，中位数仍然是 3%。

如总体样本数量是偶数，计算中间两个数值的平均值作为中位数，可以避免极值的影响。

众数的误区



如数据组庞大，较少机会出错；如数据组较小，容易有误区。

※例如，如数据组数值都在 1-100 之间，但 1 出现了 3 次，那么 1 就成为众数，虽然平均值（这种情况下比较敏感）会接近 50。

※大规模调查可以通过强调众数来操控。100 受访者对某产品的满意度在 1-10 之间打分，即使打 10 分的人数比其他分的人数只多了 1 个，10 就是众数。

- 〔视频〕算术平均数、中位数、众数之比较（国语）
- 〔参考〕算数平均数，中位数、众数

想一想〈五个整数〉

有五个整数，其平均数是 4，众数是 1，中位数是 5。求该五个整数。

解题及答案

既然众数是 1，必然最少有两个整数是 1。因为中位数是 5，第三个整数必然是 5。这个数字组是{1, 1, 5, x, y}。

如平均数是 4，五个整数的总和必然是 $4 \times 5 = 20$；即是 $1 + 1 + 5 + x + y = 20$，暗喻 $x + y = 13$。

以下说明最简单的情况：假设 x 是少于或等于 y，如 $x = y$，得出 $x + x = 13$, $2x = 13$, $x = 6.5$。明显 x 是大于或等于 5，因此 5 是少于或等于 x 少于或等于 6.5。

因此，如 $x = 5$ 就会有两个众数：1 和 5。因此可推论 $x = 6$, $y = 7$，而这五个整数必然是{1, 1, 5, 6, 7}。

数据源：http://mathschallenge.net/full/average_problem

第三章　不存在的小数字

一位统计学家建议，看到一项调查结果时就要质疑：「前后有多少个陪审团才找到这一个？」

如前所述，采用颇为偏差的样本可以产出几乎任何结果；依常规的随机采样，如规模小而又多番使用，也可以产生几乎任何结果。

「用家改用白齿牌牙膏后，蛀牙减少 23%！」仔细阅读，说明还声称调查结果来自令人放心的「独立」实验室，数据也是由特许会计师认证。还要什么更多证据？

然而，大多数人从经验中知道什么牌子的牙膏都是差不多。为何白齿牌的用家有这样的声明？这广告是否说谎？没有，况且广告不必说谎。有更简单更有效的方法。

第一个搅局的因素是样本不足，不符合统计学的要求。广告的小字说明测试组群只有十几人。[18]

有些广告会忽略这讯息，即使精通统计也只能猜想这是什么品种的诡辩。在类似的情况，十几人的的样本不是那么糟糕。几年前，有一种牙粉上市，自称「矫正龋齿相当成功。」当时的想法是该牙粉含有尿素，已由实验室证明有效。这是毫无意义的，因为这初步试验只涉及六个案例。

那么白齿牌牙膏没有说谎，又如何得出被认证的结果？让任何小组样本在半年内记录蛀牙数目，然后改用白齿牌牙膏。只有三个必然的结果：蛀牙明显更多、明显更少或没有明显变化。如果是第一或第三个情况，白齿牌牙膏把数据存盘（在看不见的地方）并重复调查。迟早，只是因为机率的操作，测试组必然出现第二种情况，值得大吹大擂，作为广告标题。无论测试组是用苏打或其他牙膏，都会出现第二种情况。

利用小组群的重要性是这样的：在大组群机率产生的任何差异很可能只是少许，不值得大书特书。减少蛀牙 2%的广告不会让牙膏大买特买。

小规模样本只凭机率产生的变化，实在不能说明什么。来一个小实验吧。

[18] 译注：许多国家的保护消费者法例要求广告说明调查的主办方，日期和样本数目。

人人都知道抛硬币花纹朝上的机率是一半一半。抛硬币十次，花纹朝上的可能有八次，这「证明」花纹朝上的机率是 80%。牙膏统计就是这样。只抛几十次，有可能得出 50%的结果，但不大可能。但是，如果耐心抛上一千次，几乎可能极为接近 50%（但不完全肯定）的结果；这才是真正的机率。要有相当数量的测试，平均规律才可以是有用的描述或预测。

多少次测试才算足够？这是棘手问题，取决于受采样调查的母体群其数量和其中差异的程度。有时，样本的数目并不是表里如一。

几年前有一个显著的例子是关于脊髓灰质炎疫苗的试验。这似乎是一个令人印象深刻的大规模医学试验：450 名儿童接种疫苗，对照组是 680 没有接种的儿童。此后不久，小区爆发流行病。曾接种疫苗的儿童没有一人感染小儿麻痹症。

但对照组的儿童也没有感染。在设计试验时，相关人员忽视或不理解麻痹性脊髓灰质炎的发病率较低。以一般发病率计算，这规模的母体群只预期有两宗病例。因此这测试从一开始就注定没有意义。测试母体群要有十五或二十五倍的规模才可以得出稍有意义的答案。

许多伟大的医学发现曾在类似的情况下急急出台。正如一位名医所说：「要赶快采用新医疗措施，以免为时过晚。」[19]

犯错的不限于医学界。公众压力和草率报导往往迫使未经证实有效的治疗提前发动，尤其面对当前庞大需求而统计数据朦胧不清。几年前的感冒疫苗和近年的抗组织胺药就是例子。这些失败的「灵药」之深受欢迎，主要是因为疾病的不可靠本质和逻辑的缺陷。感冒无需吃药，过几天就会自我治愈。

如何避免被不确定的结果愚弄？不可能人人是统计学家懂得研究原始数据。有一个很容易理解的显著性检验：究竟报告的测试数字有多大可能是真实的结果，而不是偶然产生。这是非专业人士不明白而且不存在的小数字。

如讯息来源有给出**显著水平**[20]，就更容易掌握。显著水平最简单的表达方式是机率。人口普查局给出「机率为 19/20」，表明具体的精确度。在大多数情况下，这 5%显著性水平已经够好。有一些较严格的要求 99/100 的机率，这意味着确切显著差异机率为 1%，这有时被描述为「实际肯定」[21]。

---

[19]传闻这句话出自 William Osier 爵士和 Edward Livingston。他们都同是医生和这方面的权威。
[20] degree of significance
[21] practically certain

还有另外一种可能同样有害的不存在小数字。这小数字说明事件的范围或其与平均值的偏差。平均值（无论是平均数或中位数，具体或不具体）往往流于过于简化，比无用更糟糕。一无所知通常好于一知半解；只知皮毛可能是危险的事情。

例如因为统计数据家庭有三至六人，据此规划建房，房子有两间卧室供三至四人居住。这「平均」规模的家庭实际上只是家庭总数的少数。为「平均」家庭建造房子，而忽视人数较多或较少的家庭；一些地区已经有过多两间卧室的房子，而较小和较大的单位不足。这误导而又不完善的统计已导致代价高昂的后果。公共健康小组指出：「算术平均值歪曲了实际的情况：三人和四人家庭只有 45%。35%是一人及二人家庭，20%是四人以上。」

人们面对「三至六人」的权威数字，莫名其妙地失去理智，抵消了人们从观察中得知的印象：很多小家庭，少许大家庭。

类似的不存在小数字情况是令无数父母担心的所谓「格塞我常模[22]」。家长在周刊和报章读到小孩三个月大学会坐起来，立即就想到自己的小孩。如小孩三个月大还没有坐起来，家长往往得出结论小孩是「弱智」或「不正常」等等令人反感的顾虑。由于小孩必然有一半到了三个月大不会坐起来，很多父母不开心。当然，从数学上来说，有另一半的父母发现自己的小孩「胜于他人」，他们的喜悦平衡了前一半父母的忧愁。如忧愁的父母强迫小孩符合常模，会适得其反。

这一切并不是说 Arnold Gesell 医生和他的方法有什么问题。问题出自耸人听闻或学艺不精的作家过滤了研究人员的讯息，未有留意在这过程中消失了的数字。如果这些「常模」或平均值能补上正常范围的说明就可以避免很多误解。父母看到自己的小孩是属于正常范围，不会担心那些微小而无意义的差异。几乎没有人在任何方面是完全正常，就像抛硬币一百次很难会得出五十次是花纹向上。

混淆了「正常」与「理想」让这一切变得更糟糕。Gesell 医生只是简单说明一些观察到的事实；只是担心的父母在阅读书籍和文章时以为小孩坐起来比常模慢了一天或一个月必然是比别人逊色。

对金赛性学博士的大多数愚蠢批评（其实很少人曾透彻阅读）来自把「正常」等同良好，优异，可取。金赛博士被指控把各种常见但不受认可的性行视为正常，因而荼毒青年人心灵，向他们灌输有害的思想。但他只是陈述他认为这些是正常活动；这正正是「正常」的意思，他没有加上任何「认可」的印章。他不认为他是判断这些行为是否「不可取」的权威。博士碰上了一直困扰着许多其他观察员的危险难题：提出任何情感敏感的内容而不另行草草陈述你是否支持或反对。

---

[22] Gesell's norms

不存在的小数字其欺骗性不是因为没人留意这不存在，虽然这是小数字成功的秘诀。现今对新闻工作者的批评是谴责「坐在办公室的记者」不再如老派记者去「跑新闻」，而是不加批判地重新编写政府的新闻稿。以下的不思进取新闻样本来自新闻杂志《双周刊》〈工业新发展：西屋公司冷浴法增强钢硬度三倍〉。

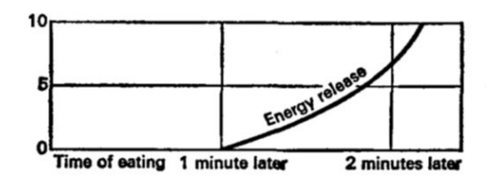
这听起来像不错的发展，直到读者试图明白这是什么意思，这句子变得难以捉摸。新浴法是否在处理后增强钢硬度三倍？抑或生产的钢铁其硬度是三倍以前的任何钢铁？冷浴法有什么作用？看来，记者只是传递文字，没有探讨其中意思，而是期望读者水过鸭背，看过了就以为快乐地学懂一些什么。这让人联想到课堂教学讲授法的旧定义：教师把教科书内容传送到学生的笔记本计算机，双方都没有动脑筋的一个过程。

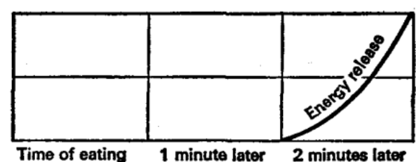
几分钟前，我寻找《时代》周刊一些关于金赛博士资料时，发现另一不堪细看的语句。这是电力公司在 1948 年的广告：「时至今日，超过四分之三的美国农场有电力可用」。这听起来很不错。这些电力公司真的很卖力。当然，小心眼的可以意译为「几乎四分之一的美国农场没有电力可用」。但是，真正的噱头是「可用」这个词语；电力公司利用这词语自说自话。明显地这并不意味着所有这些农民实际上用上电力；若然是这样，广告肯定会明确说明。所谓「可用」可能只是意味着电线挂在农场的上空或是十或百里的距离。

这是流行杂志一篇文章的标题：〈现在可以预测你的子女将来有多高〉。文章的显眼处展示一对图表：一个是男孩，一个是女孩，显示孩子成长期的身高会是最终身高的比例。「要确定孩子成长后的身高，核对现在的测量高度。」

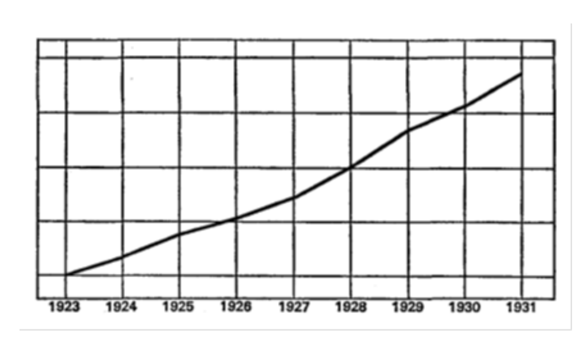
这文章和图表的致命弱点是忽略了不是所有孩子都是以同样的方式长高。有些慢慢长高后加快，有些突然长高一段时间然后趋于平稳缓慢，还有一些是相对稳定的长高。这些是基于大量测量结果的平均值。以总数或平均数计算，随机取样一百名年轻人的高度这毫无疑问是准确的，但父母感兴趣的只在某时刻的高度，这样的图表几乎是一文不值。想知道孩子将来会有多高，观察他的父母和祖父母可能得出更好的猜测。这不是很科学和准确，但至少比图表准确。

我十四岁时参加高中军训班，按身高排在矮子班，按图表我最终身高应该是 5 英呎 8 英吋。现在我是 5 英呎 11 英吋。预测身高有三英吋的错误是极为差劲的。





有两盒葡萄+坚果+麦片的早餐食品，不同的包装，都有「科学家证明这是真的！」的图表标榜「在两分钟内开始给你能量！」左图表在左边列出数字，右图省略了数字。数字没有说明代表什么，没有意思；反正两个图表都没有特别意思。图表显示陡峭的攀爬线，分别显示在进食后一分钟（左图）和两分钟（右图）后能量释放。左图的能量线爬升约快一倍，这表明绘图人员没有想到这些图表是什么意思。



这种愚蠢图表可能只是想吸引青少年或早上半梦半醒的疲惫家长。没有人会用这样的统计图来侮辱大商巨贾的智慧吧…或者会吧？《财富》杂志的广告宣传栏经常刊载某机构业务逐年上升趋势的令人印象深刻图表。图表没有数字。究竟这是业务增加一倍或一年逐年以数百万美元增加，或是以蜗牛速度每年只增加一两元，不得而知。

如平均值或图形或趋势没有包含一些重要数字，就要加倍小心。露营人士不会依赖平均温度的报告选择营地。61℃是舒适的平均温度，在加州的可选范围包括内陆沙漠和海岸离岛。但中间数忽略了范围：内陆沙漠的温度范围 15～104℃，海岸离岛是 47～87℃。

第四章　为了子虚乌有无事忙

Josiah Stamp 爵士记述 Randolph 勋爵研究收入的报告。他的私人秘书一直站在旁边。勋爵说：海关收入比去年同期增长 34%，令人欣慰。秘书纠正他，指出这只是•34%。

「这有什么区别？」勋爵问道。秘书解释 34 是•34 的一百倍，勋爵说：「我经常看见那些该死的小点，但从来不知道他们的意思。」

小数点和其他该死的差异突然出现，困扰着测试成绩的比较。不介意的话，提一个例子。国光和美莲参加智力测验。很多学生在求学时期都会参加类似测验，已成为这个时代的主要巫术偶像之一，可能要争论要花功夫才能找出测试的结果；讯息是如此深奥，经常被认为要交由心理学家和教育学家处理才是安全的。无论怎样，国光测试的智商是 98，美莲是 101。当然，智商是基于 100 的平均或「正常」水平计算。

啊！美莲是聪明的，高于平均水平；国光低于平均水平。不要纠缠于这些结论，因为任何这样的结论都是无稽之谈。

先要说清楚：无论智力测验计量的什么东西，并不是我们一般以为的智力。智力测验忽略了一些重要的事情，例如领导力和创造性的想象力，没有考虑到社交场合的判断能力，或是音乐、艺术或其他能力的倾向，更不要说努力处事和情绪平衡等性格特征。最重要的是学校最经常给出的测试是阅读测试（快速和便宜）；慢读的学生不可能拿高分。

假设我们已经认识这一切缺点，并同意智商仅仅只是计量一些定义含糊，处理抽象问题的能力。也假设国光和美莲参加的是一般认为是最好的个别测试，并且不要求任何特定的阅读能力。

智商测试声言是智力的采样。一如任何其他抽样方法的产品，智商是一个有统计误差的数字，误差影响智商数字的精确度和可靠性。

这些试题就像随机在农田采摘玉米，采摘了一百条玉米，应当对这块农田的种植状态心中有数。这样的讯息已足以和其他玉米田比较（如两块玉米田不是很相似）。如两块农田差别不大，可能要采摘更多玉米，并以一些确切的质量标准评价采摘的样本。

玉米样本能如何准确代表整块农田，可以用可能误差和标准误差[23]的数字表达。假设要在栅栏以外目测许多农田的大小，第一件事可能是先测量步行一百码的误差。如经多次步测，发现误差的平均值是三码，即是说步测有一半是超出三码，一半是少了三码。

那么能误差是每一百码有三码，或 3%，因此记录步测结果是 100±3 码。（大多数统计学家现在更喜欢用另一种但相等的标准误差[24]，只算计约三分之二的事件，而不是一半半，在数学计算方面更为方便。本书集中在可能误差，Stanford-Binet 测验也是这样使用。）

一如以上的步测例子，Stanford-Binet 智商测验的可能错误已证实为 3%。这不是关乎测验的优劣，基本上只是表达测验是否一致。所以国光的智商可以更充分地表达为 98 ± 3，美莲是 101 ± 3。

这是说国光的智商是在 95～101 的范围，他在这范围内可能是高于或低于任一智商数字，机会均等。从而可见美莲的智商高于或低于 98～104 范围任一智商数字的机会也是均等。国光智商高于 101 有 1/4 机会，美莲的智商低于 98 也是有 1/4 机会。有 3%以上机会国光不是逊色，而是优异。

这归纳为解读智商和许多其他采样结果的唯一方法是在范围之内。「正常」不是 100，而是 90～80（举例而言），也就是说比较在这范围内和在较低或较高范围的儿童才有一些意义。比较只有极小差异的数字是没有意义。必须始终记住这 ± 符号，即使（或尤其是）没有特别说明。

无视这些隐含在所有采样研究的误差，只会导致了一些极为愚蠢的行为。有杂志编辑奉读者调查为福音，主要是因为他们不理解。男读者有 40%偏爱一篇报导，只有 35%喜欢另一篇，他们要求更多类似第一篇的报导。

对杂志来说，读者的 35%和 40%之间的差异可能是重要的，但调查中的差别可能不是真实的。为了节省成本，读者样本往往减少到只有几百人，尤其是淘汰了那些谁根本不看杂志的人们。主要吸引妇女的杂志其男读者样本的数目可以是非常小。这些再细分为「阅读全部文章」，「阅读大多数文章」，「阅读一些文章」和「不看文章」各分类，那 35%的结论可能只是根据极少样本。隐藏在这些数字背后的可能误差会是如此之大，依赖这结论的编辑等同瞎子摸象。

---

[23] probable error and the standard error
[24] standard error

有时，人们为了一些数学上是真实和显著但是如此微小以至没有意义的差异而大费周折。这违背了古语的智慧：「差异如会导致差异才是差异」。一个典型例子是「老金牌」香烟为了一些子虚乌有的事情而吵吵闹闹，并从中获利。

《读者文摘》的抽烟编辑无意中开始这场闹剧。他们本来认为所有牌子的香烟都是一样的。杂志委托实验室分析几个牌子香烟的浓烟，并公布结果：全部牌子香烟的尼古丁和诸如此类东西的内容。杂志详列详尽数字，证明所有牌子的香烟实际上是相同的，抽那一个牌子没有任何区别。

你可能认为这是对卷烟制造商和构思新广告角度的广告公司是一大打击，这似乎完全推翻了香烟舒缓喉咙和对人体无害的广告声言。

但有人发现在几乎相同毒素含量的列表中，有一牌子的香烟必然排名最低；这就是「老金牌」。于是报章出现了最大标题的广告，标示这本全国通行的杂志测试所有香烟，「老金牌」含有最少数量的不良物体，但剔除了这些差异可以忽略不计的说明。最后，「老金牌」被责令终止这种误导性广告。这并没有任何影响；「老金牌」已从中得到好处。

补充材料

以会员制组织的公司讨论业积。营销部门的统计显示上月的新会员人数是全年最高。这只是部分正确。翻查记录，前两个月的退会人数也是整年最高，会员人数基本持平。上月的新会员人数也是与去年同期相若，表明这不是新趋势。[25]

[25] 数据源：http://zestsms.com/about/blog/statistically-irrelevant/

第五章　啧啧称奇的图形

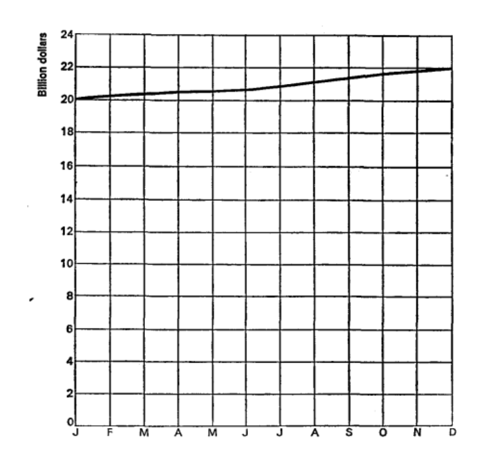
数字是恐怖的。小矮胖信心满满告诉艾丽斯，他是文字的主人；但许多人对数字没有同样的信心。也许这要回溯我们早期数学经验导致的创伤。

不管是什么原因，这对于渴望读者众多的作家，计划广告能多卖货物的公司，期望书籍或杂志大受欢迎的出版商，这确实是一个真正的问题。常见的情况是表格形式的数字是禁忌，文字又未能充份表达，往往只有一个答案：插图。

最简单的统计插图，或**图形** graph，是不同的线条，用于显示趋势很有用，实际上大家都有兴趣利用图形去知道或表达或指出或谴责或预测。

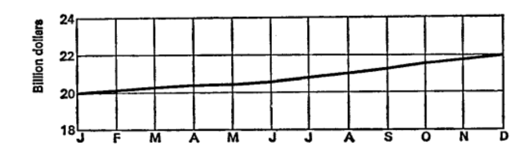
以下图形显示国民收入如何在一年之内增加 10%。

先划出方格，底线写下月份，左边标示「以十亿元计」。在方格标出数据点，连起来完成图形：



这很清楚，表明年内发生了什么，并且标明每个月的升幅。人人容易理解，因为整个图形是按比例，而且底线有 0 值作为比较。10%看来就是 10%：上升趋势是实质的但也许不是压倒性。

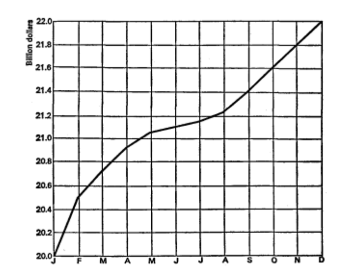
如果只是想传达讯息，这是非常好。但是，假如想赢得争论，震撼读者，促使他转化为行动，卖东西给他，这图形不够夸张。斫掉底部。



这更象样了。（也减少用纸；这是向挑剔人士反对这误导性图形的好理由。）数字相同，曲线也相同，图形也相同。没有什么是伪造的 - 除了给出的印象。匆促的读者只看到国民收入线十二个月爬升了一半的篇幅，这是因为已经不见了被裁掉的部份图形。一如语法课中的缺失句子部分，这是「不言而喻」。当然，眼睛不「理解」不存在的东西；小小的增长在视觉上成为大大的增长。

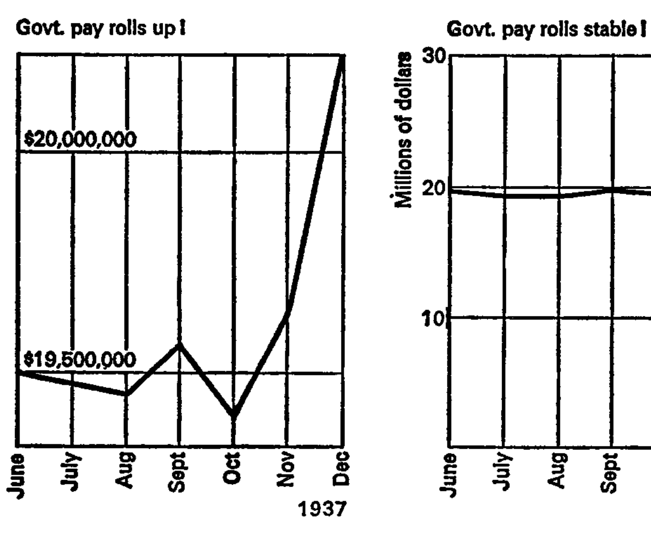
既然练习了欺骗，为什么停下来？还有进一步的伎俩可用，让微薄的 10%看起来

更活泼有力。简单地改变纵坐标和横坐标之间的比例。没有任何规则反对这样做，并且给出更漂亮的图形。要做的只是把纵坐标答比例从 2 元改写为 0.2 元。

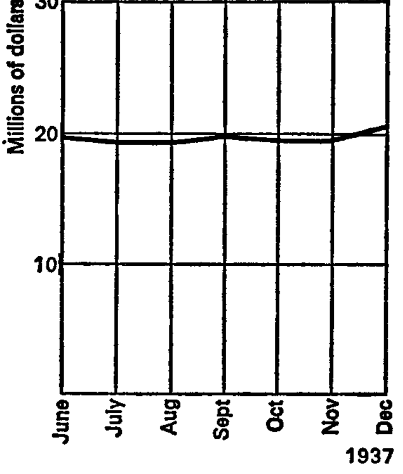


这令人印象深刻，是不是？读者会感到全国经济繁荣。这是改写「国民收入上升 10%」为「国民收入急增 10%」。这更有效，因为没有包含任何形容词或副词破坏客观性的幻想。没有人可指责你。

这样的例子不止一个。一份新闻杂志用同样方法显示股市创下新高，图形被截断，以使看起来攀升得更利害。哥伦比亚天然气公司的「我们新年度报告」的重刊图表。如果仔细阅读和分析小数字，会发现十年内生活成本上升约 60%，而天然气的成本下降了 4%。很不错，但显然哥伦比亚天然气认为还不够好，于是在 90%砍掉了图表（没有缝隙或其他警告指示）。所以，读者见到的是：生活成本增加了两倍多，天然气成本下降三分之一！



政府薪资大幅增加！



政府薪资平稳

钢铁企业曾使用类似的误导图形试图影响舆论反对工资上涨。这不是新手法，很久以前已有这样的不当行为，不仅只是在统计学专业期刊。《邓氏评论》主笔早在 1938 年看出左图的破绽：标题是「政府薪资大幅增加！」，曲线从底部急升至顶部，使得增加 4%的样子看来超过 400%。右国是修正图形：给出了相同的数字，诚实的红线仅上涨了 4%，标题改写为「政府薪资平稳」。

补充材料

图形的误区

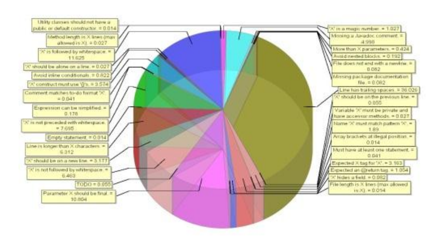
在统计学中，误导图形也称为扭曲图形，歪曲了数据，构成统计误用，导致不正确结论。

图形误导可能是因为过分复杂或制作粗糙，但精心泡制的图形也可以导致不同解释。误导性图形可能是故意，以隐瞒数据；或是无心之失：错用了绘图软件，错解数据，或是数据不适合图形表达。〔虚假〕广告特多用上误导性图形。

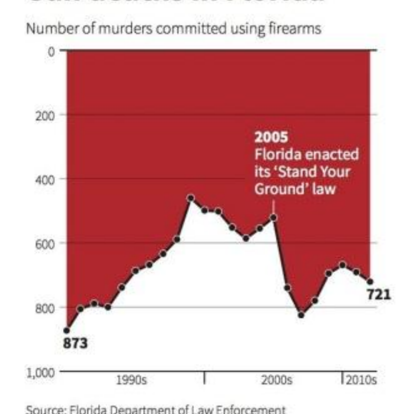
美国统计学家 Edward Tufte 创造了「垃圾图表 chartjunk」这个新字：

> 「图形的室内装修占据大量篇幅，但没有告知读音什么新的东西。装饰的目的各不相同 - 使图形看起来更加科学和严谨，使表达显得活泼，让设计师有机会展现技能。不管其原因，这些篇幅都不是数据或只是冗余数据，并且往往是 chartjunk。…Chartjunk 可以把沉闷数据变得惨不忍睹，但不能遮掩数据之不足。」[26]

不当使用图形



不需用图形而使用图形可能导致不必要的混乱 / 解释。一般情况下，图形要配上越多解释，这图形的实际需求其实越少。图形表达不总是比列表更好表达讯息。[27]



偏颇的图形

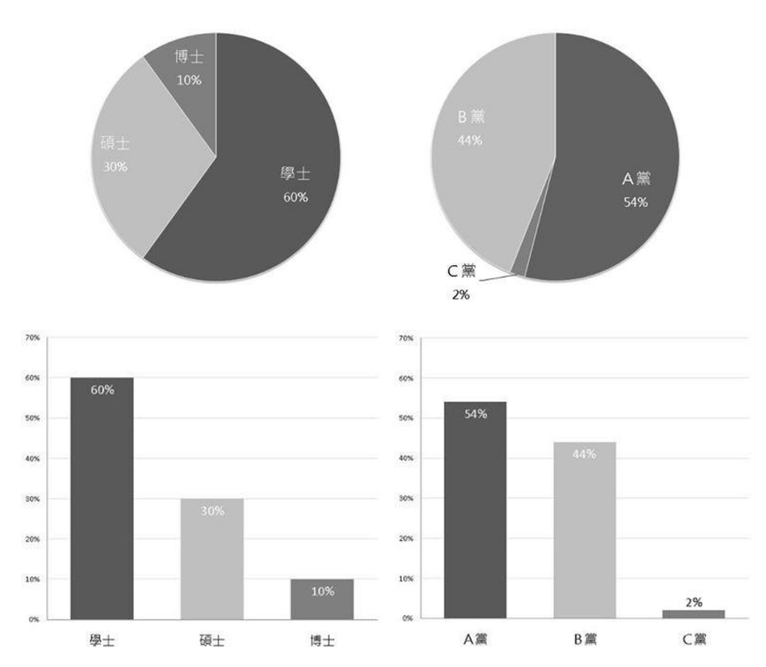
偏颇的图形标题，卷标或标题不恰当地误导读者。左图是美国佛罗里达州因枪击致死的统计图形。骤眼看来，在 2005 年订立「市民自卫法」后，枪击致死事件从高位回落。仔细一看，这图形违反一般常规，直轴是从 800 倒数至 0！数据是真实的，但严重误导。[28]

[26] *The Visual Display of Quantitative Information*.

[27] 插图取自 http://www.theusrus.de/Blog-files/pie_chart.jpg

[28] http://www.livescience.com/45083-misleading-gun-death-chart.html

饼图的误区



饼图最重要的功能在于呈现整体中各部份的组成和比例。其实条形图(bar chart)更适合比较各个组成部份的差异；虽然读者熟悉时钟角度，但还是比不上对于长度的感受。如果不看数字，条形图比较容易看出学士人数是硕士的两倍，硕士是博士的三倍。[29]

Edward Tufte 在有这样的说法：

> 「表达小的数据集，列表比图形图好很多。列表几乎总是优于愚蠢的饼图；唯一比饼图更糟糕的是几个饼图，因为读者要在多个图形之间的混乱空间要作出比较。图形图的数据密度低，又不能在视觉层面把数值排序，因此不应该使用。」[30]

---

[29] 这一段和下一段以及黑白插图取自〈饼图的使用〉，略有改写。

[30] *The Visual Display of Quantitative Information* p.178

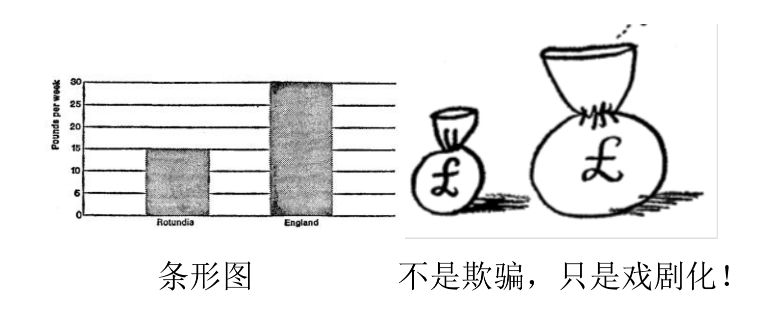
第六章　一维图形

上一代时常提到「小人物」，即是所有的人。这听起来太白鸽眼，我们成为「老百姓」。这也很快被遗忘，现在我们是「国民、公民、市民」。但「小人物」依然存在；他就是图形上的人像。

图形选择形象化，以一个小人代表一百万人，一个钱袋或一堆硬币代表一千英镑或一百万美元，一块牛排代表明年的牛肉供应；这些全是**图形统计图表**[31]，一种有用的设备，吸引注意，也能够成为流畅，狡猾和成功的骗子。

图形统计表源自普通条形图[32]，用于表达和比较两个或两个以上数据的简单和流行方法。

条形图也能够瞒骗。如图形只表达一个因素，但改变了条形的宽度和长度，或以体积难以比较的三维对象代替条形，这图形值得怀疑。被截断的条形图一如被截断的线形图同样的启人疑窦。地理书，公司声明和新闻杂志往往用上条形图，也用上吸引眼睛的图形统计图。



条形图　　　　不是欺骗，只是戏剧化！

如目的在于沟通讯息，条形图已可满足要求。但我想要更多。我想说的是英国工人的待遇远远比 Rotundian 更好，我越能戏剧化表达￡15 和￡30 的区别，我的论点越引人注目。说实话（当然我不打算这样做），我希望你从图形推断出一些东西，让你得到夸张的印象，但我不想被你看破我的招数。有一种方法，而且每天都有人这样欺骗你。

我只是画一个钱袋表示 Rotundian 的￡15，又画一个大一倍的钱袋代表英国人的￡30。这是按比例的，是不是？我追求的是你的感觉。英国工人的工资远远多于外国人。

---

[31] pictorial graph or pictograph
[32] bar chart

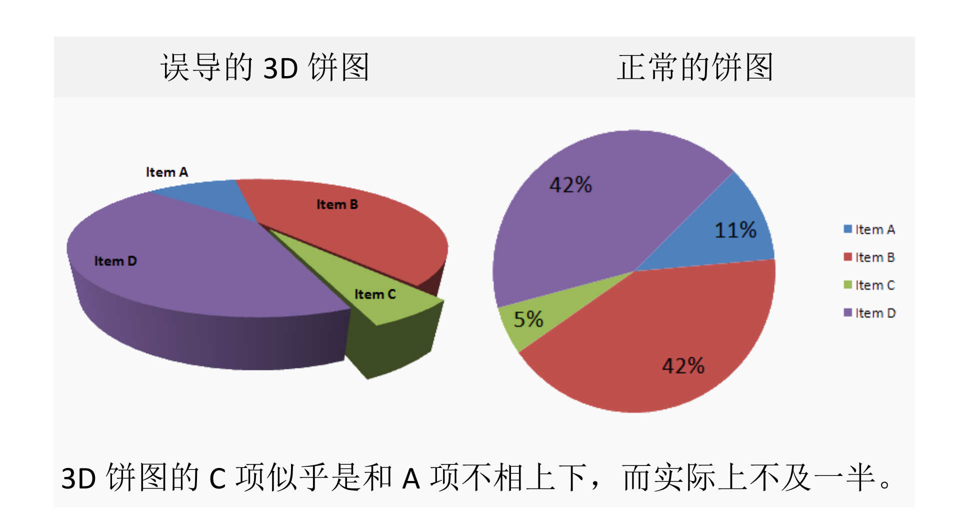
当中的诡计是这样的。因为第二个钱袋是第一个的两倍高和两倍宽，占用篇幅不是两倍，而是四倍。数字依然是二对一，但占据主导地位的视觉印象是四比一，或者更多。因为这些三维图像是立体的，第二个钱袋的厚度必然是第一个的两倍。几何教科书指出类似立体的体积随着任何维度的立方而改变：2x2x2=8。如第一个钱袋有￡15，第二个应有￡120。

那确实是这巧妙小图给出的印象。虽然是说「两倍」，我实际留下了八比一压倒性比例的持久印象。

你也很难指责我我有任何犯罪意图。我只是随波逐流。新闻杂志反复这样做，一如上例的钱袋。

补充材料

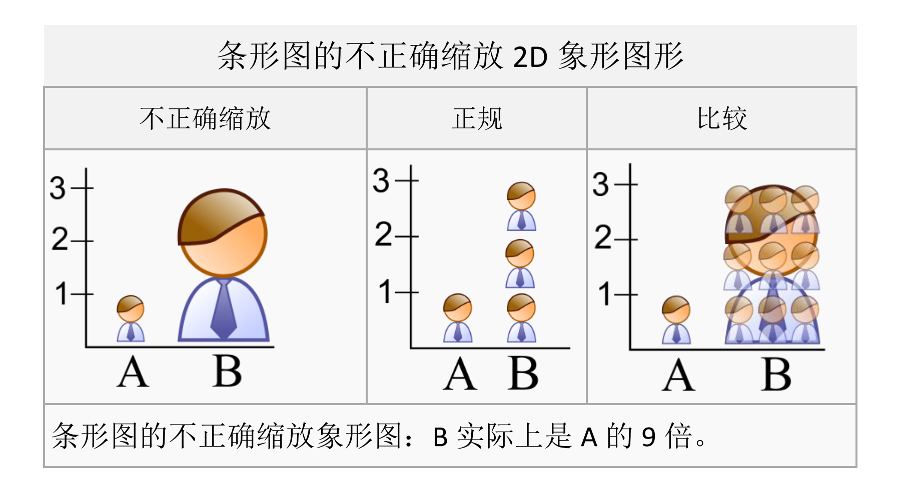
很多统计图形不适合三维(3D)形式，饼图特别如此。由于消失点效果，即使同样大小，3D 饼图靠近读者的部份会看起来比较大块，较远的部份比较小。这扭曲了数据的呈现。只是为了美观而牺牲精准表达，说不过去。下面的例子说明这现象：



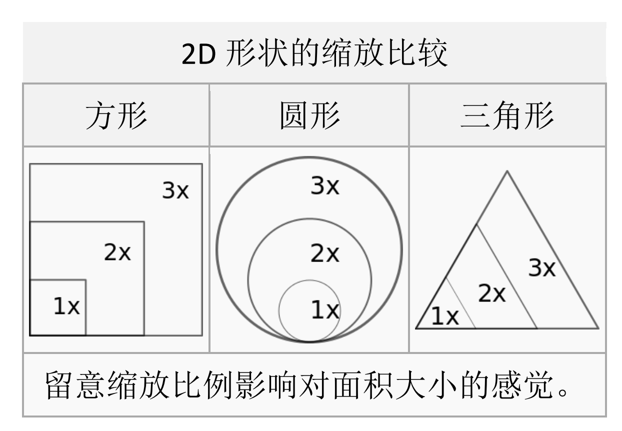
3D 饼图的 C 项似乎是和 A 项不相上下，而实际上不及一半。

不正确的缩放

条形图使用象形比例，不应均匀缩放，因为这导致误导性比较。读者看到的是象形图的面积，而不是高度或宽度，导致比例以平方面积解读。



条形图的不正确缩放象形图：B 实际上是 A 的 9 倍。



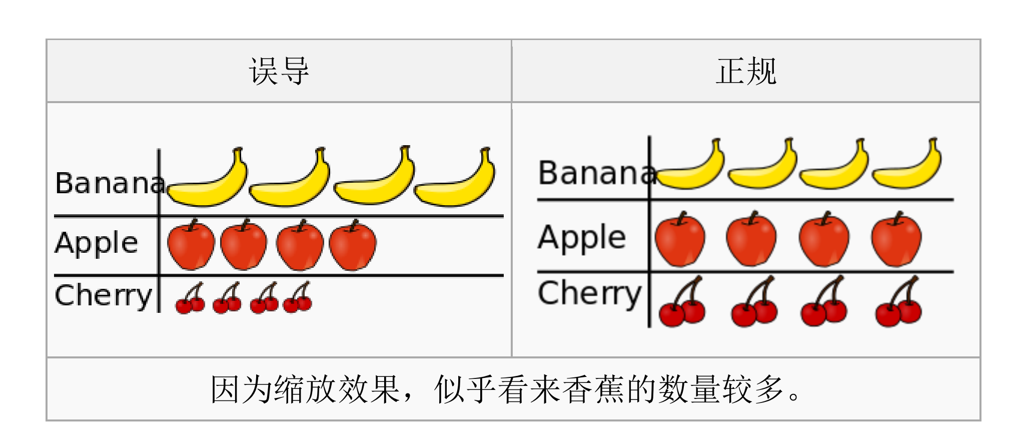
留意缩放比例影响对面积大小的感觉。

3D 象形图不当缩放导致立方效果。



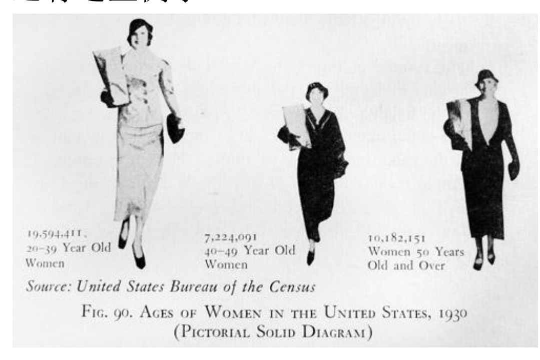
这 3D 象形图显示 2001 年房屋销售比去年有增长。因为没有直轴说明，读者无法理解变化；两倍的缩放看来是八倍(2^3)。

不当缩放的 3D 象形图误导读者以为项目实际上改变了大小。

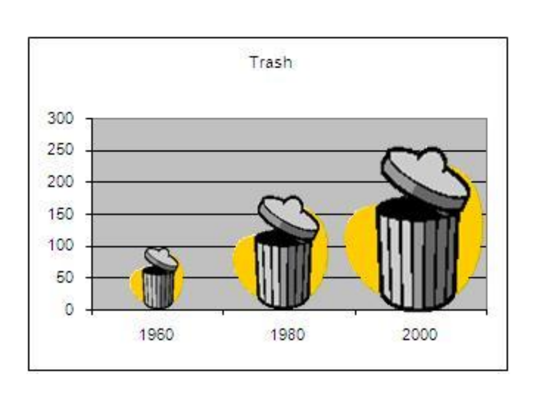


因为缩放效果，似乎看来香蕉的数量较多。

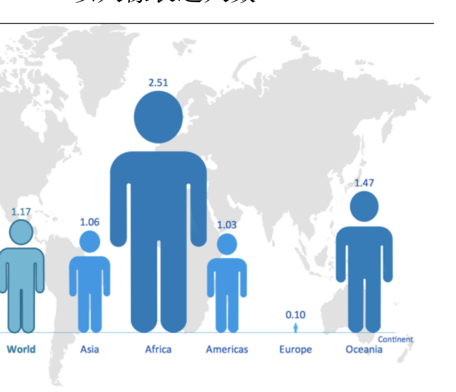
还有这些例子：



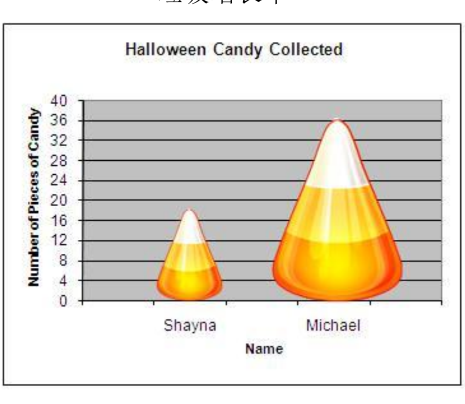
以人像表达人数[33]



垃圾增长率[34]



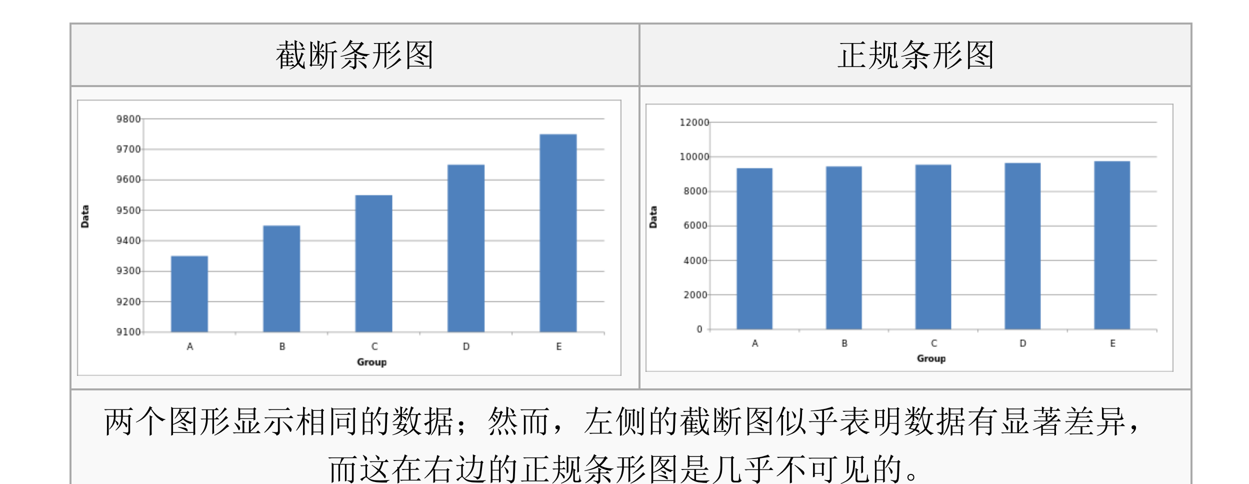
人形表达[35]



几多倍？[36]

截断图形 truncated graph（也称为撕裂图 torn graph）的直轴（y 轴）不是从 0 开始，可用于显示微小的变化或节省空间，但可能导致把少许变化错认为重要变化的错误印象。如数值是在狭窄范围，有些软件（如 MS Excel）其默认功能会自动制作截断图形。

截断条形图	正规条形图



两个图形显示相同的数据；然而，左侧的截断图似乎表明数据有显著差异，而这在右边的正规条形图是几乎不可见的。

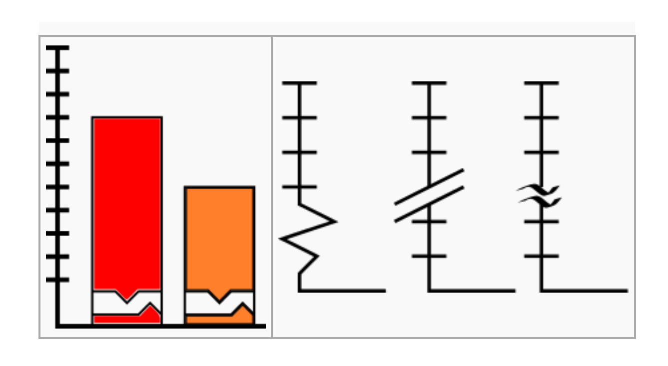
[33] http://www.timwallace.info/b/wp-content/uploads/2011/03/womendiagram.jpg

[34] http://yale.edu/ynhti/curriculum/images/2008/6/08.06.06.03.jpg

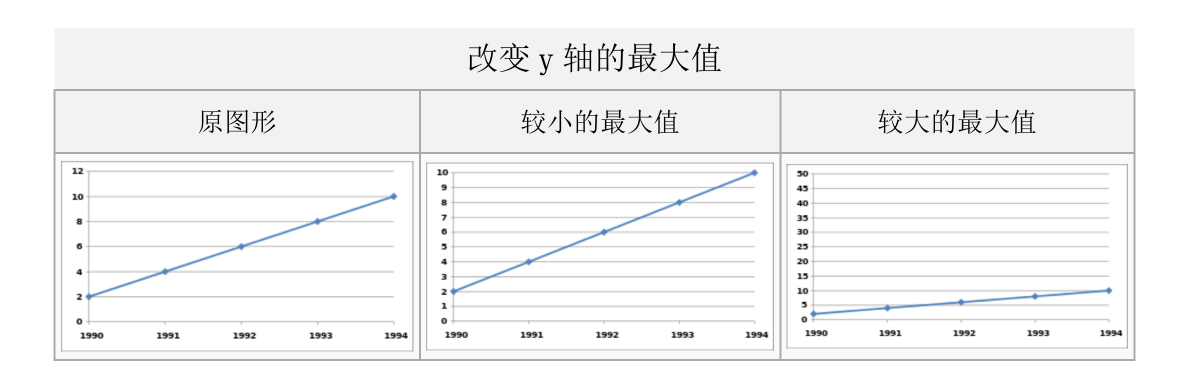
[35] http://www.conceptdraw.com/solution-park/resource/images/solutions/picture-graphs/GRAPHS-AND-CHARTS-Picture-graphs-Population-growth-by-continent-Sample.png

[36] http://yale.edu/ynhti/curriculum/images/2008/6/08.06.06.11.jpg

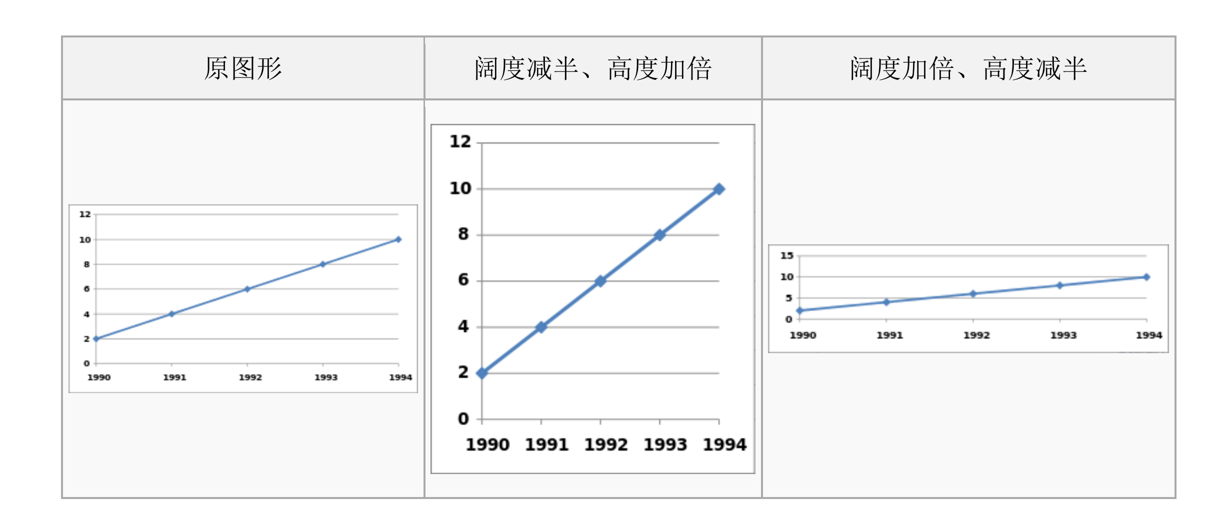
应适当提醒读者直轴被截断。



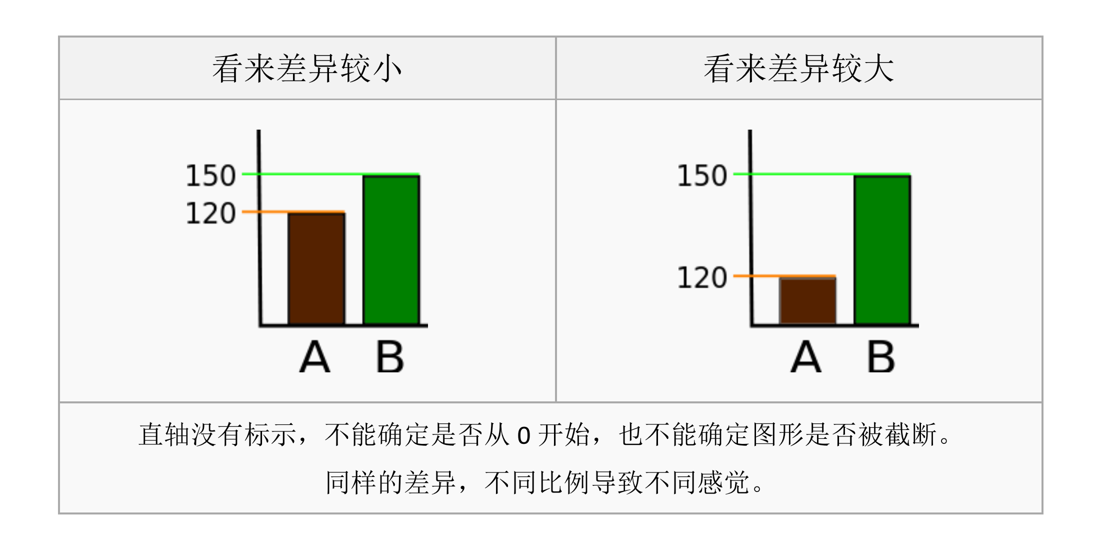
改变直轴的最大数值会导致不同的感觉。



改变图形长阔比例会导致不同的感觉。

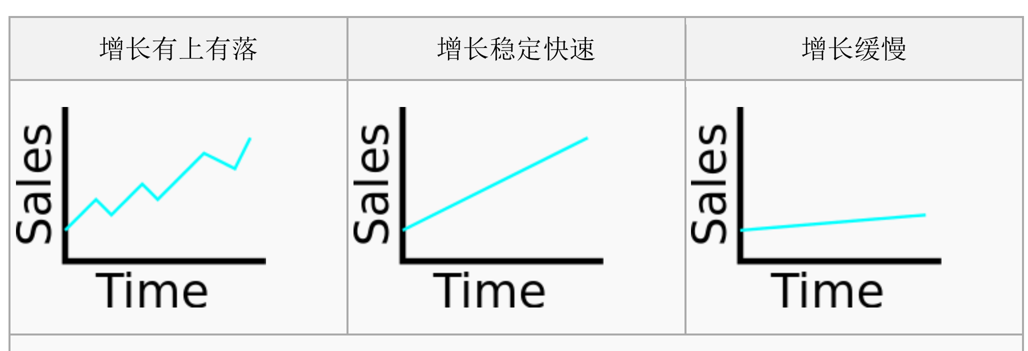


没有比例的图形往往用于夸大或减轻项目差异的感觉。



直轴没有标示，不能确定是否从 0 开始，也不能确定图形是否被截断。
同样的差异，不同比例导致不同感觉。

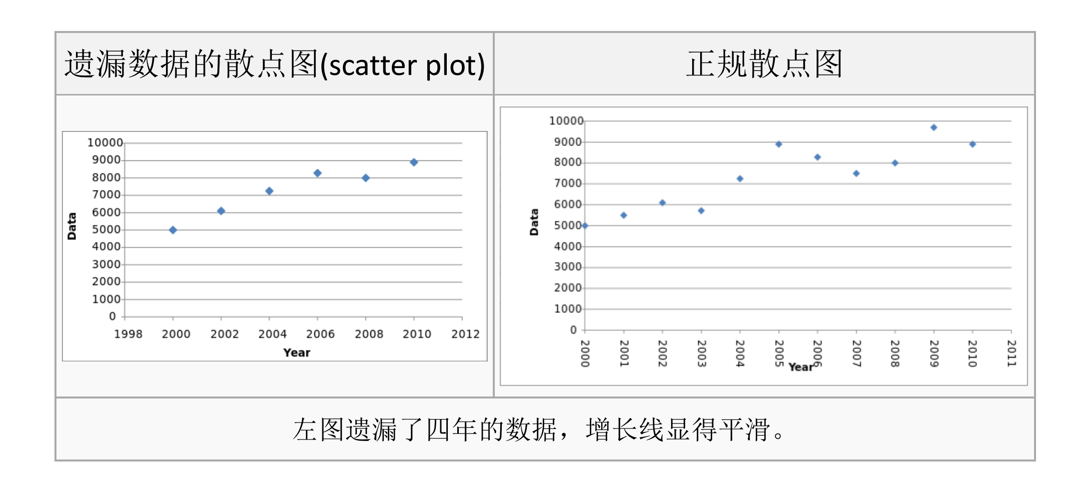
另一例子：



这三个图形表达同样数据，因而数据的实际(x, y)斜率是一样的，
但因为各自的直轴比例不同（没有标示），所以视觉上有不同的斜率。受操控的误导。

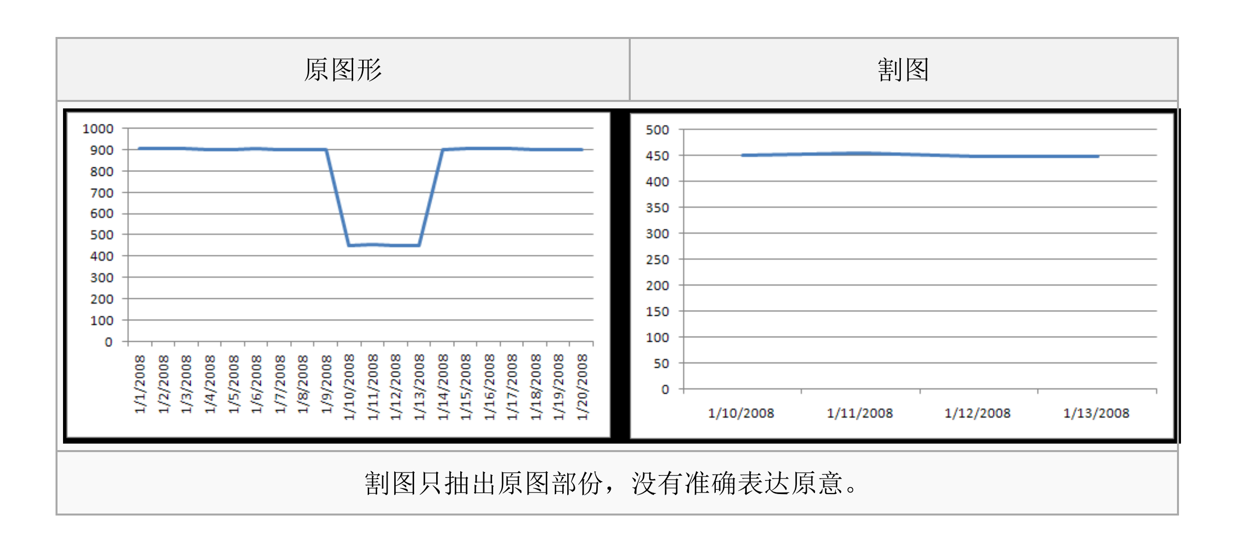
数据遗漏

遗漏了数据的图形就是误导的图形，不能从中得出正确结论。

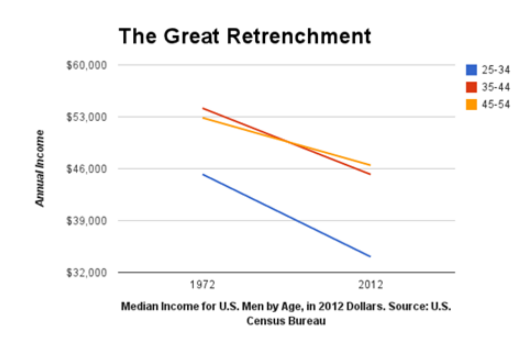
遗漏数据的散点图(scatter plot)	正规散点图
	
左图遗漏了四年的数据，增长线显得平滑。	

不正当的割图

从其他图形抽出部份为割图，应保留（有时强调）原来的特征。

原图形	割图
	
割图只抽出原图部份，没有准确表达原意。	

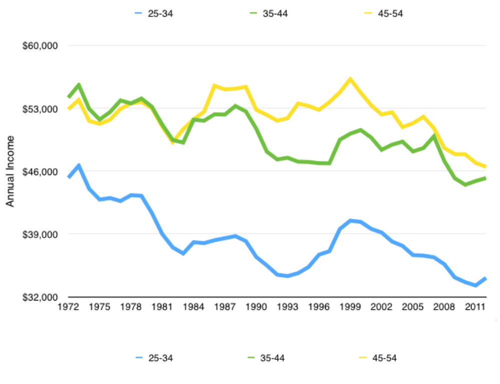
剪裁数据和扭曲图形



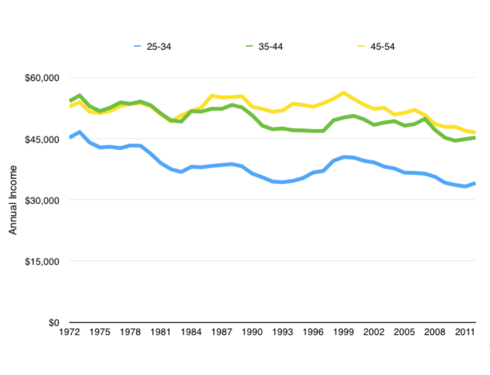
2013 年，彭博通讯社企业及市场编辑发表署名文章〈美国男士四十年来收入下降〉For U.S. Men, 40 Years of Falling Income，附上插图说明三个年龄组群的美国男士的中位数收入下降，下降斜率颇为惊人。文章集中讨论 1972 年和 2012 两年的数据。

数据来自美国人口调查局，彭博是有声誉的通讯社，作者不是初出茅庐的见习记者，报导应该是可信的吧？

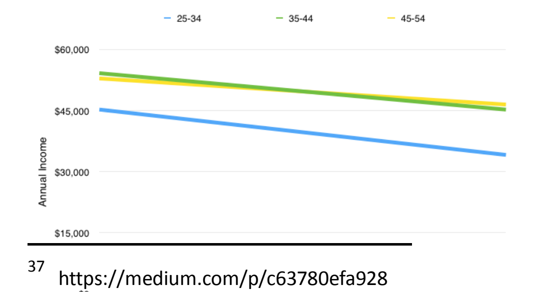
Eric Portelance[37]留意到这截断图（直轴不是从 0 开始）问题多多，于是深入研究相关数据，发现原作者只集中讨论 1972 年和 2012 年的数据，似乎故意忽视了在这期间的多年数据。



重新制作的没有截断的连续图给出不同年份的数据，得出不同印象。总体而言，中位数收入依然呈现下降趋势，但斜率不是第一图的剧烈。45-54 岁组群是相当稳定，直至 2000 年才有下降。



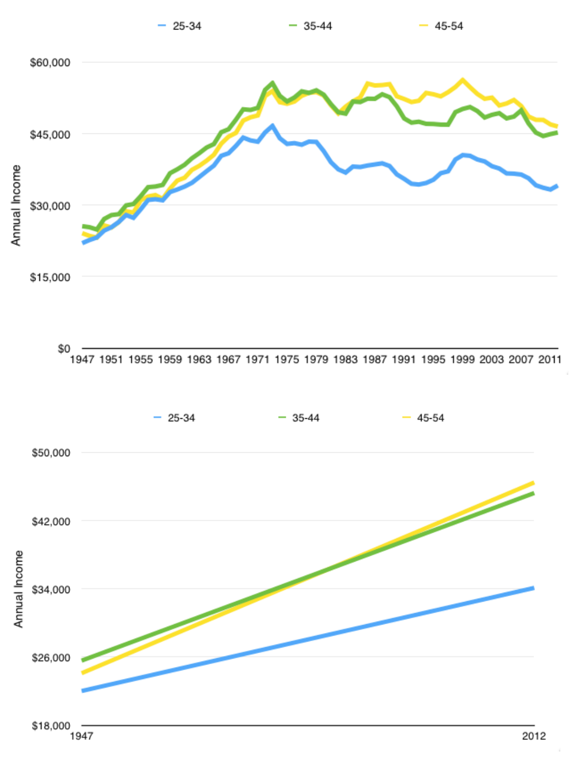
若是图形没有截断，回归正规从 0 开始，中位数下降的斜率可说是缓慢。



若原图没有截断，中位数下降的斜率不是文章强调的「危险」。

[37] https://medium.com/p/c63780efa928

Portelance 进一步找出人口调查局的全部数据，发现彭博编辑「忽略了」1947 至 1972 年的趋势。



1947 至 2011 年的全部数据得出不同的结论：收入持续上升，直至 1971 年见顶，之后有些年龄组群保持平稳，有些逐年下降。研究主题应该是「为何如此？」而不是「美国男士四十年来收入下降」。

如追随彭博作者只选用两年的数据作为起点和终点，不同的选择（只选 1947 和 2012 年）得出完全不同的结论！

这是统计谎言的典型例子。

第七章　半吊子的数字

一名印度法官忠告热心的年轻英国公务员：「当你年纪大一点，就不会热衷于统计数据。印度非常热衷于积累统计：收集，添加，提高至 n 次幂，取立方根，并准备精彩的图形；但绝不能忘记的是这些数字每一个都是来自村长，他们喜欢说什么数字就说什么！」

如果不能证明你想证明的什么，证明别的东西，假装是同一东西。人们面对统计统计数据的冲击时会发呆，几乎不会注意到其中的差别。半吊子的数字是非常有用的手段。

药厂不能证明新药能治感冒，但可以大字发布实验室报告：半公克新药在试管内 11 秒杀死 31,108 枚病菌。要确保实验室是有信誉或有令人印象深刻的名字。拍摄穿白袍的医生拿着报告。

但不要提出几个噱头：在试管中有良好效用的药剂可能不会在人的喉咙有作用，不要说明杀死什么病菌以免混淆。谁知道是什么病菌引起感冒，特别病源可能不是病菌？事实上，没有人知道试管中各种细菌和感冒有什么关连，但人们不会深入理解，尤其是感冒病人。

也许，这例子太明显了，人们多了对感冒的认识，虽然广告页面从来少不了这些声东击西的例子。

在种族歧视的年代，奉命调查以「证明」不是这回事，这是艰巨任务。你可以计划一次民意调查，或更好的是委托有声誉的机构调查；向有代表性的母体群发问：黑人的就业机会是否和白人一样？每隔一段时间进行一次调查，最后得出趋势的结论。

普林斯顿大学民意调查中心曾经调查这题目，发现得出的民意表里不一。每位受访者除了回答主题问题，还要回答其他问题以测试他是否歧视黑人。调查发现种族歧视观念最严重的受访者，对就业问题的答案往往是正面。同情黑人受访者有三分之二认为黑人就业机会逊于白人；有种族歧视观念的人有三分之二认为黑人就业机会不逊于白人。明显这项调查对黑人公平就业机会说不清是什么情况，反而揭露了人们看待种族的另一面。

因此，在种族歧视的年代，调查黑人的公平就业机会，会得出「黑人就业没有问

题」的结论。情况越差，这些半吊子数据让调查看来更好一些。

「执业医生有 27%选择金叶牌香烟，多于任何其他牌子。」暂由不论这说法是否虚假，只要问这说法有什么问题。大多数人的反应可能是：「那又怎样？」医学界受到尊重，但医生知道香烟品牌的讯息是否多于普通烟民？他们是否有特别知识选择危害最小的香烟？当然他们不是这样。然而，「执业医生有 27%选择金叶牌香烟」似乎意味着更多的什么。

「实验室试验证明大力牌电动榨汁机功能提高 26%。」。这听起来真不错；直至真相揭露是大力牌电动榨汁机的功能是与老式手动榨汁机比较。大力牌电动榨汁机可能是市场上功能最差的，那个 26%数字是完全不相干。

不是只有广告客户玩弄数字，更多的是从数字中导出没有关连的结论。一篇交通安全的文章报导：「晚间七时的交通意外是早上七时的四倍」，因此在早上开车更安全。数据没有问题，但结论不可靠。晚上的交通比早上繁忙，所以较多意外，与文章的结论没有关系。

如果没有留意这些数字是半吊子的数据，你可以被任何交通工具事故的统计数据吓得半死。

相比 1910 年，更多人死于飞机意外。现代的飞机是否更危险？废话。现在的飞机乘客是以前的数百倍，仅此而已。

「据报导，去年的铁路意外死亡人数为 4,712 人。」这很吓人。真相是有一半死亡人数是因为汽车司机闯红灯，在道口与火车相撞，其余大部份是跳车的霸王乘客，只有 132 人是火车乘客。甚至这数字也没有很大比较意义，除非这连接到总乘客里程。

知道火车，飞机或汽车去年的意外伤亡数字，也要同时知道每百万乘客—公里数字，才可以知道风险比率。

声东击西有很多法宝，一般手法是并列两种看来相关或相似，但其实没有关连的项目。某企业与工会的关系恶劣，人事部经理受命「调查」员工对工会的投诉，必然可以找到一些相关投诉，理直气壮声称「员工有 78%反对工会」；实情只是搜集一些不经分类的投诉和埋怨，汇集为另一套数据。这没有证明什么，但似乎是完成了调查。

当然，这是双面刃；工会也可以随时「调查」，「证明」员工对企业的诸多不满。

企业的财务报告多的是这些半吊子数字。留意出乎意料的庞大利润和隐藏在某他名目的利润。汽车工人工会有这样的报导：

「公司公报去年利润三千五百万元，占销售额的 1.5%」，少得可怜。换一个三毛钱的灯泡已耗上二十元销售额。员工甚至想到要节省用纸。公报的利润当然不是全部利润，其余的隐藏在折旧，特别折旧和储备。

同样要留意百分比。通用汽车公报本年九个月的税后销售利润增加 125%，投资部门盈利增加 448%。这究竟是好是坏？视乎你的观点。

同样，读者来函为 A&P 商店辩护：「商店每千美元销售额只赚了十元，不应被谴责为奸商。」咋听之下，这样的利润确实微不足道；住房抵押贷款和银行贷款的息率在 6%之上。A&P 公司结束超市业务，把资金存入银行赚取利息岂不是更有生意头脑？

心法在于投资年回报率不是等于销售总额的利润。正如另一位读者投函解释：「如每天早上以$0.99 买货，当天以$1 价格售出，利润只有 1%，但全年 365 天的投资盈利是 365%。」

任何数字都有许多表达的方式。例如，可称之为销售回报率 1%，投资回报 15%，一千万美元的利润，利润比去年增加 40%，或比去年下降 60%，方法是选择一个最适合当前目的的数字，希望没有几个人会理解这是如何不完善反映了情况。

不是所有半吊子数字是故意欺骗的产品。许多统计数据，包括对大家非常重要的医疗数据，是因为源头失真而被扭曲。一些微妙事项如堕胎，婚外生育和梅毒都有惊人的矛盾数据。美国最近公布的流感和肺炎数字，奇怪的结论是这些疾病几乎都局限在南部三个州，占报告病例约 80%。实情是这三个州依然把流感和肺炎列为必须申报的病例，其他州已经停止申报。

一些关于疟疾的数字没有意义。1940 年前，美国南部每年有数十万例，现在只有极少数，似乎短短几年内有极大改进。实情是现在只呈报确诊为疟疾的病例，而之前是包括了南方人惯称的感冒或发冷。

1898 年的美西战争，海军死亡率是 9‰，同一时期的纽约市平民死亡率是 16‰。海军征兵人员后来用这些数字来宣传在美国当海军更安全。假设这些数字是准确的，看看这两个数字为何几乎毫无意义。两个组群没有可比性。美海军主要身体健康的年轻人；纽约市平民包括婴幼儿，老人和病人，他们全都有较高的死亡率。

两个数字不能证明符合海军标准的士兵活得更长寿，但也不能反证。

在发明脊髓灰质炎疫苗之前，沮丧的消息是小儿麻痹症是史上最严重，当年比以往任何时候都更多病例。

专家检视这些数字，发现几件令人鼓舞的事情。其中之一是当年的小儿数目是破纪录的数字，如发病率不变，病例数字也会水涨船高。另一发展是父母更多认识脊髓灰质炎，即使轻症病例更愿意求医就诊。最后是有了财政诱因：有更多的小儿麻痹症保险和慈善组织的更多援助。所有这一切令人怀疑小儿麻痹症达到新高的说法，后来的死亡总人数证实了怀疑是合理的。

值得留意的事实是死亡率或死亡人数往往比发病率或发病人数是更好的衡量 - 仅仅是因为报告和记录死亡率或死亡人数是较为尽心和准确。

美国每四年就有一次半吊子数字的热潮。数字没有周期，而是四年一度的选举来了。共和党在 1948 年 10 月发表的竞选声明完全是建立在似乎是互相关连但原来互不相关的数字：

> 1942 年，当 Dewey 当选州长时，一些地区老师的的最低工资低至每年$900。今天，纽约州学校的老师享有世上最高的薪水。Dewey 州长接纳他委任的委员会调查结果，在 1947 年提取部份盈余实时增加教师薪金。因此，纽约市教师的最低薪金是$2,500-5,325。

完全可能 Dewey 先生是教师之友，但数字不是这样说话。这是比较「之前」和「之后」的老把戏，从$900 急增至$2,500-5,325 听起来是极大改进，但没有说明$900 是农村地区教师的最低工资，而$2,500-5,325 只是纽约市的范围。Dewey 州长可能改善了教师的待遇，也可能没有。

之前和之后的比较照片是杂志和广告的熟悉特技。拍摄两次，告诉你新油漆涂层可以做到什么区别。在两次摄影之间，客厅已经添加新家具，有时「之前」的照片只是很小，光线不好的黑白照，「之后」版本是全彩色大照片。比对照片显示模特儿用护发素的前后对比：天哪，她确实好看得多，但仔细检查会发现大部分的变化是因为她的微笑，光亮头发。是摄影师的功劳，不是护发素。

补充材料



2007 年，英国的广告声称：「多于 80%牙医推荐高露洁牙膏」。一般人从广告得出的印象是 80%牙医推荐高露洁牙膏，余下的 20%推荐其他牌子。

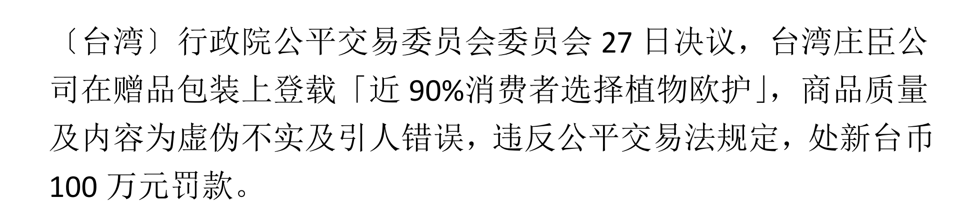
英国广告标准局介入调查，发现数据来自高露洁赞助的市场调查（但没有公布），而且受访牙医可以推荐多款牙膏，不是只选一项。调查数据显示至少有另一牌子和高露洁的得分不分上下。

英国广告标准局下令禁制广告。[38]



2009-10 年，体育用品公司 Reebok 声称 EasyTone 和 RunTone 跑步鞋经实验室测试，「证明只需穿上跑步鞋走路，比一般跑步鞋有助强化腿筋和小腿 11%，臀部肌肉更高达 28%！」

美国联邦贸易委员会调查发现这完全没有科学根据，被判罚款二千五百万美元。[39]



〔台湾〕行政院公平交易委员会委员会 27 日决议，台湾庄臣公司在赠品包装上登载「近 90%消费者选择植物欧护」，商品质量及内容为虚伪不实及引人错误，违反公平交易法规定，处新台币 100 万元罚款。

中央社报导，公平会表示，台湾庄臣依据博舆市场研究顾问于 2006 年 7 月间进行的市场问卷调查，在其赠品包装广告上宣称，近九成消费者「选择」植物欧护。

公平会指出，但经调查，该问卷其实是将庄臣的欧护植物防蚊液与另一品牌防蚊液，进行清爽不油腻偏好的比较，而非购买的比较，广告却未批注「九成」的比较基础，恐致消费者误导。

[38]数据源：http://www.telegraph.co.uk/news/uknews/1539715/Colgate-gets-the-brush-off-for-misleading-ads.html
[39]数据源：http://www.investopedia.com/financial-edge/0612/4-examples-of-misleading-health-ads.aspx

公平会表示，此外，该问卷调查以随机抽样方式进行，就 100 位受试者现场使用两种产品后调查，姑且不论样本数是否足以支持该广告宣称内容，广告宣称「近九成消费者选择欧护」，显然与问卷调查结果有别，因此认定广告不实。[40]



Centrum 在 1997 年的广告声称「十个美国人有九个未能从食物摄取所需的营养素，缺少了重要的维生素和矿物质。」该声明引用 1976 至 1980 年间进行的一项调查，发现在调查当天，受访者只有 9%记得要进食水果和蔬菜的每日推荐量，因此得出结论高达 91%的美国人缺少维生素（可能包括你！）。

这说法问题多多：（一）这不能证明那些人缺少维生素；事实上，他们可能在前一天已进食足够数量的水果和蔬菜；（二）只是一天的饮食不足以计量整体饮食习惯。食物摄入量应以几星期计算；（三）即使摄入数量低于推荐量也可以有充足营养。[41]



Vioxx 是一种非甾体抗炎药，类似阿司匹林或布洛芬。Merck 药厂的直销广告耗资亿万美元（2000 年花费了 1.6 亿美元）。该药物于 1999 年被 FDA 批准，直至 2004 年才停用。这是源于一宗法律诉讼声称该药物引起 23,800 宗心血管病例（包括心脏病发作），跟进研究发现服用 Vioxx 的患者其心血管病例统计上显著高于安慰剂患者。

这种不安全药物如何得到 FDA 批准推出市场。事因原有研究发表时，药厂排除了三宗心肌梗塞的病例，从而改变了统计显著性。可以想象药厂雇用的科学在重重压力下「忘记」这三个病例，或是他们不理解统计显著性的意义。[42]



1995 年，英国药物安全委员会向十九万名医护人员发出警告：第三代口服避孕丸增加了在腿部或肺部形成血

[40] http://dasanlin888.pixnet.net/blog/post/34467926
[41] 数据源：http://www.statisticshowto.com/misleading-statistics-examples/
[42] http://www.statisticshowto.com/how-significant-is-significant-the-vioxx-scandal/

块，有潜在的双倍致命风险。这警告导致在 1996 年有一万三千宗堕胎手术。所谓「潜在的双倍致命风险」原来是基于以下的数据：每十万名服用第二代口服避孕丸的妇女有十五人患上可致命的血块；服用第三代口服避孕丸的则增至二十五人。作为参照，没有服用避孕丸的妇女每十万人有五宗病例。是的，风险是增加了，但比怀孕的风险要小得多，不值得那么令人震惊。[43]



统计师被医生告知她的乳房 X 线检查呈阳性反应，她询问医生她患癌的机率是多少？。医生给出令人震惊的答案：80%。她遍查文献，找到正确答案是 10%，更令她震惊的是许多医生给出不同答案：20%医生回答 10%、20%医生回答 1%、60 %医生回答 81 或 90%。

不是医生看不懂数字，而是有太多研究报告被断章取义，渲染夸大。[44]

---

[43] http://news.bbc.co.uk/2/hi/health/313848.stm
[44] http://www.statisticshowto.com/even-physicians-dont-understand-statistics/

第八章　「后此谬误[45]」又来了

要估算荷兰或丹麦的家庭生了多少孩子，你可以乱猜，或是计数他们房子屋顶的鹳巢。[46]

统计术语描述鹳和新生儿两者之间有「正相关关系」，有 A 就有 B。

这个古老神话实际说明更有价值的意义：容易记住和提醒我们两个因素之间的关联不足以证明在前的 A 引起了其后的 B。

在鹳和婴儿的例子，很容易找到与两者相关的第三个因素：大家庭住在大房子，大房子有更多烟囱让鹳鸟筑巢。

但在其他情况，不总是那么容易发现因果关系的假设缺陷，尤其是流行偏见认为这是有特别意义。

有人研究和证实烟民的大学成绩是低于非吸烟者。很多人很高兴，这说法流传到现在。这样看来，要有好成绩是在于放弃吸烟；再进一步的结论是吸烟让人变蠢。

我相信这项研究是正确完成：有诚实和精心挑选的足够样本，相关性高等等。

其中的谬误颇为古老，经常出现在统计材料，躲在可观的数字之下。谬误就是：因为先有 A，后有 B，所以 A 导致 B。既然吸烟和学业不走在一起，因此吸烟导致学业不佳。但也可以倒转来说：学生成绩不佳驱使他吸烟草，但不酗酒；这结论也可以证明是对的，也得到证据的支持。但这不能满足宣传手法。

更好的结论是两者没有关连，两者都是第三因素的产物。是否喜欢交际的学生较少时间看书而多抽烟？或者之前某人证实外向性格与成绩低落之间有相关，这关系比成绩与智力之间关系更为明显？也许外向性格比内向的人更多抽烟。问题的关键是有很多合理解释，很难只是坚持己见只挑选一个。但很多人是这样。

为了避免掉落「后此谬误」的谬论作出错误判断，你需要仔细检查任何关乎「彼此相关」的陈述。这种谬误有几种类型。

[45] Post Hoc 一个事件发生在另一事件之前，并不一定是后者的原因，也译为「事后谬误」。
[46] 图片取自 http://www.todayifoundout.com/wp-content/uploads/2013/05/stork-340x400.jpg。欧洲民间传说鹳是送子鸟。

一种是偶然产生的相关性。你可搜集一组数字来证明一些不太可能的事情；但如再试一次，可能无法证明。一如「牙膏防止蛀牙」的广告，你只需扔掉不想要的结果，广泛发布那些合心意的结果。如只是小样本，很有可能发现你想得到一对一事件之间的一些实质性关联。

常见的一种共变是其中的关系是真实的，但不可能确定那个变量是「因」，那个是「果」。在某些情况下，因果关系可能会时不时改变从属地位，或两者可能同时是「因」也同时是「果」。人们的收入和持有股票之间的相关性可能是这样。有更多钱就多买股票；有更多股票，收入越多；说不准是那一个导致另一个。

也许最棘手的是变量互不影响，但有真正的相关性。这方面有颇多研究，例如烟民的学业成绩差劲；有太多医学统计虽然证实相关关系是真实的，但这「因 A 而 B」的关系只是猜测而矣。作为废话或伪相关的统计例子，有人兴高采烈地指出：马萨诸塞州长老会牧师的薪金和古巴甜酒价格有密切关系。

何者为「因」？何者为「果」？换句话说，长老是否受益于或支持甜酒贸易？这太牵强了，明显是荒谬之言。紧记世事多的是「后此谬误」，只是更为微妙隐蔽。长老和甜酒的例子很容易看到这两个数字齐齐增长，是因为第三因素的影响：世上万物的价格都在增长。

〔欧洲〕人们提到六月的自杀率最高，也提到最多人在六月结婚。是否自杀驱使较多人结婚？或是较多求婚不遂的人自杀？稍微更有说服力（但同样未经证实）的解释是在整个冬天舔着抑郁伤口的人本以为到了春天会雨过天晴，可是六月来了，他仍然感到绝望，…。

要注意的另一个结论：推断得出的相关性已超越引以为证的数据。很容易表明多雨水，玉米和农作物生长得更高更好。似乎雨水是好事。但连绵数月的强降水会损坏甚至破坏农作物。正相关关系只能维持到某一点，然后好事变坏事。超过一定的雨量，下雨越多，玉米收成越少。

当然，「相关性」的倾向经常不是被描述为一对一的理想关系。高个子男生的体重超过矮子男生，这是正相关关系。但是可以很容易找到一个六英尺的高个子体重及不上五英尺的矮子，所以相关性是小于 1。负相关简单说明「此消彼长」：变量 A 增加，变量 B 会下降。在物理学这是「反比」：灯泡的光线越远越弱。这些物理关系往往有完美的相关性，但是企业或社会学或医学数字很少是如此整齐。即使学历一般与收入成正比，但往往有许多反证。请记住，相关性可能是真实和基于真实因果关系，但如在单一事件中确定任何行动，可能是几乎一文不值。

有无数研究证实大专以上学历与未来收入挂钩，大学派发无数小册子吸引学生。我不否定这意图，我赞成教育，特别是课程包括《统计学入门》。这些数字已经明确证明拥有大学学位的人赚更多。当然，有很多例外情况，但趋势是强劲和明确的。

唯一的错误是有人利用这些数字和事实得出完全没有根据的结论。这是后此谬误的最佳例子。有人认为这些数字表明：如果你上大学，在这三、四年间你可能赚到的收入是高于以其他方式消磨这三、四年。这种没有根据的结论其依据是基于同样毫无根据的假设：因为曾受大学教育的人赚更多钱，是因为他们上过大学。其实我们不肯定知道：这些人即使没有上大学，可能都会赚更多。一些事实强烈表明正是如此。大学学生有两个群组多得不成比例：富家子弟和聪明学子。聪明的人即使没有上大学，可能已经有很好的赚钱能力。谈到富家子弟…钱生钱有多种方式。无论是否上大学，富家子弟很少落在低入息阶层。

销量庞大的星期日报刊有以下这段对话，也许你会觉得有趣，因为同一作家有另一篇文章〈流行观念：对或错〉。

> 问：上大学对你终生不结婚的机会有什么影响？
> 答：如果是女生，一生老处女的机会挺高。男生刚好相反，很少终生不娶。

美国康奈尔大学调查 1,500 名典型的中年大学毕业生，发现男生有 93%已成婚（相对于一般人口只有 83%）。但中年女性毕业生只有 65%结了婚。大学毕业生中的老处女是一般人口终生不嫁妇女的三倍。

十七岁的小美看到报导，知道如果她去上大学，婚姻大事的前景很不乐观。而且统计资料的来源颇有声誉。是的，报导有引用康奈尔大学的统计数据，但结论不是仓促读者所认为是来自校方的。

这又是案例：利用真正的相关性强加诸未经证实的因果关系。也许这一切是倒过来说。即使这些女生没有上过大学，仍然会终生不嫁，比例甚至可能高于大学女生。如果这说法的可能性并不优于作家坚持的结论，这也许也是猜测而矣。

事实上，有证据表明有终生不嫁倾向的女士更有可能上大学。金赛性学博士似乎找到了性欲和教育有一定相关性，而性状可能在大学预科年龄期已形成。这更令人质疑上大学会影响人们结婚的说法。

所以，小美注意：这是未必如此。

医学文章曾经提出严重警告，指出喝牛奶的人患癌的机会增高。在美国新英格兰，明尼苏达州，威斯康星州和瑞士这些大量生产和饮用牛奶的地方，癌症似乎变得普遍，而在牛奶稀缺的亚洲国家斯理兰卡罕见癌症。文章也指出美国南方各州少喝牛奶，癌症病例也较少。此外，有人指出经常喝牛奶的英国妇女患上某些类型癌症是少喝牛奶的日本妇女的十八倍。

只要稍为深入研究这些数字就可以得出不同解释。癌症主要是中年或以后的疾病。瑞士和前文提到的国家同样的是国民长寿。在那项英日妇女研究，英国妇女比日本妇女平均年长十二年。

Helen M. Walker 教授提出证明，解释有趣但愚蠢的说法；证明假设每当两件事情一起变化必然有因果关系的谬误。调查妇女的年龄和某些物理特征之间的关系，可以计算步行时脚的角度，会发现老年妇女的角度往往较大。可能实时反应这反映因为脚的角度加大，所以她们长老了。人人都看出这是荒谬的解释。似乎是年龄增长导致脚的角度增大；大多数妇女长老了，脚的角度加大。

任何这样的结论很可能是虚假和必然是不合情理。要适当得出正确结论，研究应在一段时间内观察同一妇女或类似组群。这会消除一个可能的因素：老年妇女成长时，被教导走路时脚要朝外，而现在的年轻少女被教导这样的姿势不正确。

如有人（通常是有利害相关的人）对某项相关关系大做文章，首先看看这是否这类型的关系：产生于事件流程，时间趋势。我们这时代很容易发掘到任何两项事物的正相关关系：大学学生人数，精神病人数目，香烟消耗量，心脏病数字，使用 X 光机次数，加州学校教师的薪俸等等。认为其中一些事物是另一些事物的「因」显然是愚蠢无理。但太阳之下无新事，每天都有人提出。

以统计学方法和迷惑的数字和小数点来阐释因果关系，只是比迷信好一点，但往往比误导更严重。新赫布里特群岛的岛民一直相信体虱是健康良好的表征。他们观察了几百年，目睹身体健康的人通常有体虱，而生病的人往往没有。观察本身是准确和有见识；历久以来，这些非正式的观察往往都是。从证据中得到这些原始结论：体虱让人健康，人人都应该有体虱。对此，我们很难有什么说法。

正如上文指出，在统计磨房处理比这还要稀少的数据，直至常识的目光再也无法穿透，已经为医疗界和许多杂志和专业医学期刊赚钱不少。精明观察者终于弄清楚新赫布里特群岛的现象。事实证明，几乎每个岛民大部分时间都有体虱；可说是正常状态。然而，当病人发热（很可能是由那些体虱传染），病人体温变得太热，体虱离开这不再舒适的居所。这案例的因果完全混淆、扭曲、扭转和混在一

起。

补充材料

错误的因果关系

当统计测试展示 A 和 B 之间的关系，通常有五种可能性：

1. A（因）导致 B（果）。
2. B（因）导致 A（果）。
3. A 和 B（因）互相导致对方出现（果）。
4. A 和 B（因）一起导致 C（果）。
5. 观察得的关系纯属偶然（没有因果关系）。

第五个可能性可透过统计测试来量化，计算出来的机率与前四个可能关系发生的机率一样大，但事实上应变量之间是没有关系。

如调查发现沙滩泳客购买雪糕的人数与遇溺人数有相同趋向，没有人会断言雪糕导致遇溺，因为这是明显地无关。遇溺和购买雪糕的人数明显与第三个因素（沙滩上的人数）相关。

但这谬误的例子不是笑话：例子是「接触化学品 X 会导致癌症」的诸多报导。把「接触化学品 X 的人数」代替「购买雪糕的人数」；把「患上癌症的人数」代替「遇溺的人数」。在这情况下，即使两者没有真正的因果关系，但统计上依然有关联。例如，如某地方有「危险」（即使并不危险）的化工厂，中产家庭因恐惧而迁离，诱使更多低收入家庭搬到该地。然后发现低收入家庭患上癌症的数字上升，于是推论化工厂是元凶；其实这可能是基于较差的膳食和生活环境或是较低档次的医疗服务。

第九章　统计误世

通过使用统计材料以误导他人，可称为统计操控，或是「统计误世[47]」。

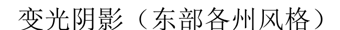
本书的书名和一些内文似乎暗示所有这些操作都是意图欺骗的产物。美国统计协会的分会会长曾为此斥骂我。他说：大多数不是欺骗，而是无能。他的说话有意思，但我不能肯定统计学家认为那一项批评更为不恭敬。可能更重要的是要记住：扭曲统计数据及其操作并不总是专业统计人员所为。统计学家的成果被推销员，公关专家，记者，或广告文案扭曲，夸张，过度简化，或通过选择扭捏。

但无论在任何情况下谁是有罪的一方，很难说这是无心之失。杂志和报纸经常夸大炒作虚假的图表，很少减斤扣两。在我的经验，业界提出的统计参数很少报大报喜，往往是表达差于数据。另一方面，少见工会聘请无能的统计人员做出数据差于表达的统计。

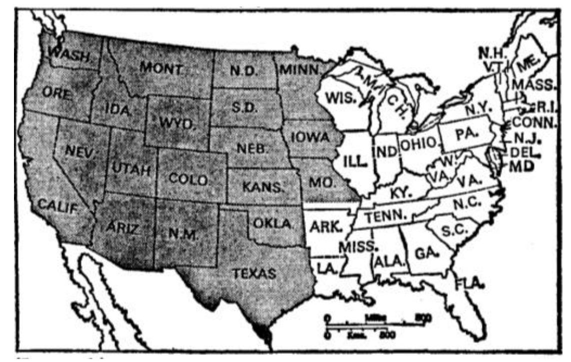
只要这些错误是一面倒，很难归结于笨拙或意外。

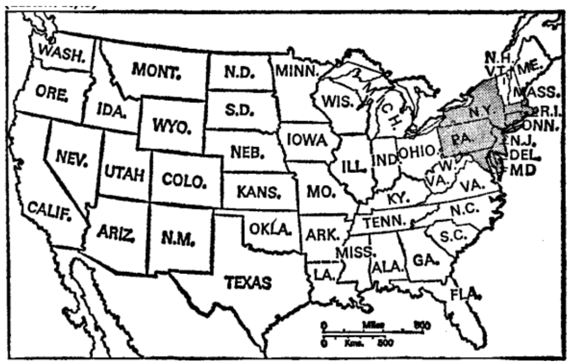
歪曲统计数据巧妙手法是利用地图。地图隐含许多变量，其中事实可以被掩饰，关系被扭曲。我最喜欢的「变光阴影[48]」奖杯颁发给不久前波士顿第一国民银行发表和转载极广，包括所谓纳税人群体，报纸和《新闻周刊》。

变光阴影（西部各州风格）



变光阴影（东部各州风格）





为了表示我没有作弊，地图加了 MD, DEL 和 RI。

该图显示目前联邦政府拿走和花费的美国收入部份，利用有色部份表示密西西比河以西各州（除了路易斯安那州，阿肯色州和密苏里州部分），其联邦政府支出

47 statisticulation

48 The Darkening Shadow

已等于各州国民的总收入。

欺骗谎言在于选择地广人稀的各州，其收入相对较少。以同样的诚信（和同样的不诚信），绘图者可能已开始在纽约或新英格兰着色，得出极为更小但更令人印象深刻的阴影。使用相同数据，他可以给出产生完全不同印象的地图，但没有人有兴趣发表。至少，我不知道有任何强大群体有感兴趣发表偏少的公共开支。

如果绘图者目标只是传达讯息，很容易做到。他可以选择一组中间状态的州份，其总面积与总收入占国民收入比例相同。

这张地图公然误导，不是宣传的新把戏，而是经典手法。同一家银行不久前公布显示联邦政府在 1929 年和 1937 年开支的地图版本，很快被辑录为「可怕插图」歪曲事实的例子。这间银行依然故我发表绘图，而更有见识的《新闻周刊》和其他人一直照搬可也，没有警告也没有道歉。

如果你认为现在有通货膨胀，看看这个。有一段时间，美国人口普查局想出了在年报陈述「平均家庭收入为$3,100」。但同时报章又报导 Russell Sage 基金会给出的同样数据是可观的$5,004。也许你高兴知道大家生活得不错，但也可能感受到这数字与你观察所得不符。也许你认识的人不是基金会认识的群组。

人口普查局和基金会的数字怎会如此不同？普查局是说「中位数」，也是应该如此；但即使基金会是说「平均数」，差别也不应该如此巨大。基金会解释数据来自把美国人民个人总收入除以 149,000,000，得出人均$1,251；四口之家即共有收入$5,004。

这样奇怪的统计操控有两方面的夸大：（一）使用「平均数」而不是较小和更多讯息的「中位数」（上文有讨论）；（二）假设家庭收入是家人数目成正比。我有四个孩子，也希望事情是这样，但事实不是。四人家庭的收入绝对不是两人家庭的两倍。

公平地说，基金会的统计学家可能不是存心欺骗，应该说他是想表达人们捐献而不是受惠的意思。有趣的家庭收入数字只是副产品，但这欺骗行为已广泛传播；这是不能轻信平均数的最好例子。

表面精确会赋予最声名狼藉的统计数据看来有斤两。考虑小数点的例子。调查一百人昨晚睡了多少小时，比如说得出总数为 7,831 小时。首先，任何这样的数据远远不可能精确。大多数人的的猜测有十五分钟或更长时间的错误，而且不能保证这些错误〔在数据集〕会自我平衡。有人失眠五晚，只记得折腾了半晚。无论

如何，调查算出各人的平均睡眠时间为 7.831 小时，听来你是知道自己在做什么。如果发表的数字是 7.8（或近乎 8）小时，这不是什么惊人的吧。这是拙劣的接近数值，比几乎任何人的随意猜测都没有什么启发性。

马克思以同样手法制造精密的虚假氛围。他要计算工厂的「剩余价值率[49]」，开始汇集一些假设、猜测和整数：「假设废品为 6%…。成本为整数 342 英镑。有一万个纱锭…假设成本为 1 英镑。折旧率假设为 10%。假设工厂租金为 300 英镑。这些数据是由一位曼彻斯特市纺纱工人提供，可以信赖。」马克思利用这些近似数值算出剩余价值率是 6%。[50]

百分比是制造混乱的沃土。一如令人印象深刻的小数点，百分比为不精确数据罩上精密的光环。美国劳工部曾表示华盛顿特区的兼职家庭在指定月份领取的交通津贴，有 49%是每星期 18 美元。细查之下，这个百分比原来出自两个只有四十一项优惠的案例。基于少数案例的任何百分比都可能误导；直接给出数字更能提供更多讯息。如百分比带上小数点，小心欺诈。

「现在购买圣诞礼物，节省 100%！」。这广告听来像是圣诞老人自掏腰包，但只是制造混乱。原来是减价 50%。节省 100%是指新价格的 100%；这是事实，但不是广告吹嘘的事实。

标准石油公司的文献走得更远：「割价 14～220%」。这似乎要求卖方支付买方一笔可观费用去拉走油腻腻的东西。

某公司宣布货品销售获利 3,800%，算自成本 1.75 元和售价 40 元。计算利润百分比有多种方法（必须说明）。如果以成本计算，利润率是 2,185%；以售价计算是 95-6%。这间公司发明了新方法，得出了夸张的数字；而这似乎常常发生。

甚至纽约时报转载美联社报导时，也犯了「移动基数[51]」的错误：「经济萧条今天狠狠地打了工人一记重拳。印第安纳波利斯建筑贸易工会属下的管道工，泥水匠，木匠和其他工获得工资增加 5%。这只是他们去年削减工资 20 %的四分之一。」

表面看来这算法很合理；但跌幅是基于一个基数（工人之前的工资），而今年的加薪是基于另一个较小的基数（现有薪酬水平）。

小小心算即可指出以上是统计误算。为简单起见，假定原来工资是每小时$1，削减 20%即是下跌到$0.8。$0.8 增加 5%即为$0.04，这是削减额的 1/5，不是 1/4。

---

[49] rate of surplus-value

[50] 看不清原文的计算方式，笼统译之。

[51] Shifting Base

一如许多诚实谎言，这篇报导夸大了一个本来很好的故事。

这一切说明：要抵消减薪 50%，下一次加薪必须争取 100%。

「转移基数」做成许多折扣的错觉。「五折再八折」不是原价的三折，而是四折，因为「八折」是以较小的「五折价」为基数。

一种装模作样的欺骗手法是把不对号但似乎相关的东西相加。一代又一代顽童都用这一套证明他们不用上学。

你可能还记得吧。一年 365 天，减去在床上度过的 122 天（三分之一），再减去饮食时间 45 天（每天三小时）。剩余的 198 天要扣了 90 天暑假和其他假期 21 天。剩下来的日子甚至不够分配给周末。

你可能认为大企业不会利用这古老和明显的伎俩，但美国汽车工会坚持汽车企业依然用这一套来对付他们。

每一次罢工期间都会出现这谎言：汽车企业声称罢工每天的损失是若干百万美元。这数字来自如罢工工人全力工作会制造出来的汽车，加上供货商的损失。一切可能的被加进来，包括销售商的损失。

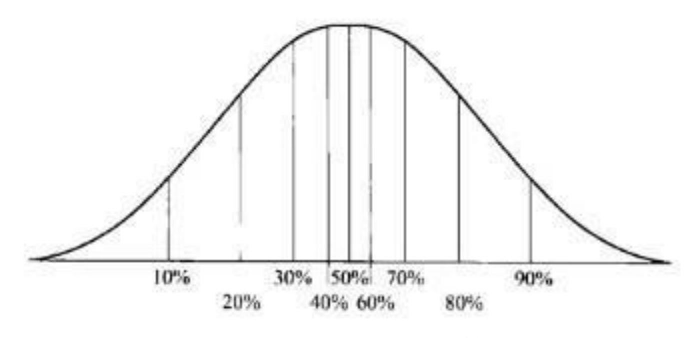
同样奇怪的概念是百分比可以自由加在一起。《纽约时报》书评版这样说：书价和作者收入之间的差距越来越大，是由于生产和材料成本大幅上升。在过去十年，厂房及制造费用上升多达 10-12%，材料上升 6-9%，销售及广告开支向上攀升超出 10%。只是一间出版社，这些林林总总加起来至少有 33%；较小规模的出版社几近 40%。

其实，如果每个成本项目上涨约 10%，总成本必然也以 10%同样比重攀升。把各项成本的增加叠加起来，是鬼话连篇。今天你买了二十种日常用品，发现每种都比去年价格上涨 5%，会否有人大声疾呼：「生活成本增加了一倍！」

这就像路边小贩解释他的兔子三明治如何能卖得这么便宜。「我必须渗一些马肉：一只兔子的肉渗入一匹马的肉。」

工会反对一位「聪明笨伯」老板定义每小时平均工资：正常工时每小时$1.5，加班每小时 $2.25，周末加班每小时$3，共三小时得出平均每小时工资$2.25。这有意思吗？

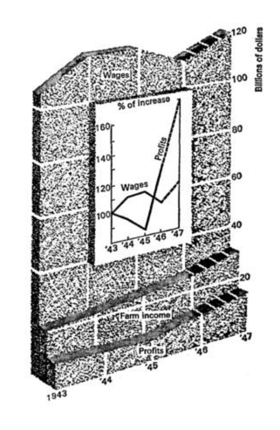
混淆「百分比 percentage」和「百分点 percentage point」是容易堕入的陷阱。如投资的利润从去年的 3%攀升至今年的 6%，可以低调只是「增加三个百分点」，或是大事张扬「增加了百分之百」。特别是民意调查最常利用这种手法。



正态分布的百分位数

百分位数 percentile 是统计术语，容易骗人。这基本上是将一组数据从小到大排序，并计算相应的累计百分位，某百分位所对应数据的值就称为这百分位的百分位数。例如代数班有三百名学生，按各人成绩排序，百分位数 99 是成绩最佳前三名，其后三位是百分位数 98 ，依此类推。百分位数有奇怪而容易混淆的地方：百分位数 99 的三位学生的成绩远远优秀于百分位数 90 的三位，而在百分位数 40 至 60 的几十位学生成绩可能几乎相等。这是由于世事万物的正态排序惯常呈钟形曲线：最优最劣只占少数，大多数趋向中位值。

偶尔统计人员发动内战，旁观者察觉到事有蹊跷。美国钢铁工会为了争取改善待遇，指出以 1939 年为基数，行业的生产力已大大提高，所以钢铁企业有能力加薪。工会没有说明因为特别事故，1939 年的产量超低。企业的欺骗手法也不甘示弱，坚持员工的总薪资已有上升。这不是平均时薪，而是全体员工的总收入，其中包括许多早期以散工身份加入企业，后来转为长工的人员；即使工资水平没有上升，这么多任务人的收入必然会增加。



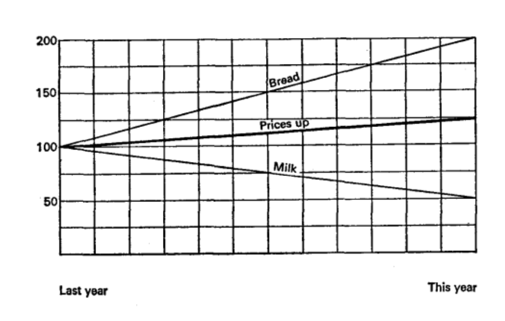
《时代》杂志的图形一向精益求精。这张插图说明图表可以是百宝袋，任由劳方资方随意抽出所需的证据。这插图其实是表达同样数据的两张插图迭加一起。

方格图显示工资和利润（以十亿美元为网格线比例），很明显两者都上升，而去年工资的增长是利润的两倍。以美元计，工资增长是利润的六倍。巨大的通胀压力似乎是来自工资。

白底插图显示工资和利润增加的百分比。工资线相对平稳，利润线大幅度向上。由此可见通胀压力主要来自利润。

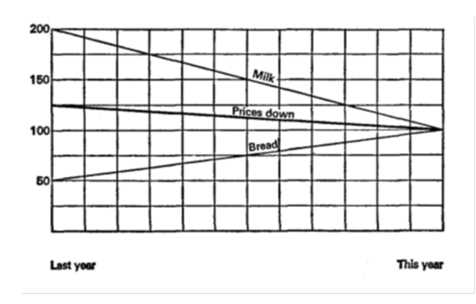
你可以得出自己的结论，或是更好的看到任何一方都不是通胀的主因。能够及时简单地指出争论的主题不是表面的非黑即白，已经有助人们理解。

指数数字[52]至关重要，影响百万受薪族的工资。因此要提醒各位这也是任人剪裁的。



以最简单的例子为例：去年，牛奶每瓶 10 便士，面包每个也是 10 便士。今年牛奶降价到 5 便士，面包是 20 便士。这说明什么？生活成本是涨了还是降了？还是没有变化？

考虑以去年为基期，把当时价格作为 100%。由于牛奶价格减半(-50%)而面包价格翻了一倍(+200%)；50 和 200 的平均值为 125，价格涨了 25%。



再试一次，以今年为基期。牛奶本来是现价的 200%，面包是现价的 50%。去年价是今年的 125%。

为了证明成本水平没有改变，简单切换为几何平均值[53]，并以两个年份为基准。这少许有别于算术平均值，但也是完全合法，并在某些情况下是最有用和启发。要得到三个数字的几何平均值：各数相乘，得出立方根。四个数字取第四根，两个数字取平方根。就是这样。

以去年为基准，价格水平为 100。实际是每项乘以 100%，取其平方根，得出 100。以今年为基准，牛奶是去年价格 50%，面包是 200%，200 乘以 50 得出 10,000；其平方根 100 即是几何平均值。各项价格没有上涨或下跌。

事实是尽管统计有数学基础，但既是艺术，也是科学。在这范围内有许多操作，甚至扭曲。通常情况下，统计学家必须选择表达事实的方法，这是主观的过程。在商业现实中，他不太可能选择对己不利的方法，一如广告撰稿人不会描绘赞助商的产品不坚实和不够档次，他会说轻巧和经济。

即使是学术界可能也有偏差（可能无意识），特别想证明某这一点。

[52] Index number
[53] geometric average

这表明我们要三思统计材料，在报纸和书籍，杂志和广告的事实和数据。但随意拒绝统计方法也是没有意义。这就像拒绝阅读，因为作家有时用文字来掩饰事实和关系，而不是披露公开。

补充材料

数据集的误区

大量的数据才能得出有效的平均值，并准确预测趋势。一万人的数据优于一百人。只有 3-5 个数值的数据集，得出的结果并不真实。

数据集不仅要很大规模，也要很广泛。地质学家调查沙漠数据，在沙漠十个不同地点收集 100 个数据，要比在同一地点收集 1,000 个数据更准确。

有两个人，有一位双腿截断了。无论选择哪一种平均值，只要不被看出只有两个样本，那么就无法辩驳「人平均有一只脚」的结论。

有些调查故意这样做。例如，人口统计想要找出男性更倾向某种职业，那么只需要调查男性人群。

一些小项调查经常错误地把控制集的调查结果等同普遍结果划等号。小项调查没有办法调查广泛、随机的城市人口，学院调查经常方便地面向大学生人群，尤其是心理学测试实验。即使调查报告说明情况，但新闻机构为了发表耸人听闻的报道，往往把细节模糊，利用院校层次的调查结果来以偏概全。

使用不平衡的数据集撒谎的做法非常狡猾。技巧是把那些其实并不能相提并论的数据放在一起比较。例如，十万人口的新城镇在十年新增一万人，比较原本只有十个居民的小村落在十年增多十个人，那么就可以理直气壮地总结小村落人口增长更快。

有时市场调查会利用这技巧来发表销售数据。调查苹果和橘子的销量，但是调查到了一半橘子由于存货不足卖光了，但调查依然继续，那么苹果销量就会远远高于橘子，即使苹果并不是真的更受欢迎。

解读调查数据的误区

许多事物的因果关系涉及多个甚至无数的因素，调查往往不能孤立少数因素以设

计对照组研究。

另一方面，这些复杂关系又方便了调查从中撮出一些有利本身观点的结论。常见的统计陷阱是调查测试包含大量应变项(dependent variable)，方便找出一个有利自己的似是而非的因果关系。

第十章　如何反驳统计的谎言

最后一章解释如何看透虚假的统计，如何从中找出可信可用的统计。

不是所有眼见的统计讯息可以诉诸化学分析或踏实研究的诚实测试。以下五个简单问题有助找出答案，避免受骗。

（一）谁的统计？

要寻找的第一个答案是偏见：进行调查和发表结果的一方有什么动机？实验室是为了理论，名声还是收费而去证实什么？报章是否追求销路？劳资双方是否要鼓吹某个工资水平？

留意故意的偏误。这可能是直接的错误陈述，可能是模棱两可的不明确声明，可能是选择有利数据和忽略不利数据，转换测量单位（例如选择有利的数据作比较），可能选用不适合的计量单位（例如采用平均数，而中位数能披露较翔实或更多讯息），以没有说明的平均数挂羊头卖狗肉。

公司宣布 3,003 人持有公司股票，平均持有 660 股。这是真实的数据，但没有说明三位大股东已持有总股票数量四份之三，另外三千人共持有余下的四份之一。

要留意无意的偏见，这往往是更危险。在 1928 年，许多统计学家和经济学家发布图表和预测，证明经济繁荣，无视经济结构中的裂纹。

面对这些「证据」，至少要一看再看是谁发表这些统计数据，无论是声名显赫的政界、科学实验室、甚至大学。报导引述：「某某大学研究发现…」，要注意的不是「某某大学研究发现…」，而是谁在引述，因为引述的结论往往是作者之言，不一定是某某大学的结论。

《芝加哥商业期刊》大事公告该期刊调查 169 间企业有关对抬高价格和囤积居奇的结果：三分之二企业宣布他们面对远东地区的加价，是一如既往由企业吸收消化部份。期刊说（每遇上这些说话，要加倍留神!)：「调查显示这些美国企业没有追随他人提价。」这是明显的要质疑：「是谁这么说？」由于期刊可被视为有利害关系，这也顺延到第二个测试问题：

（二）他怎么知道？

取样

期刊相当取巧：事实是调查对象为 1,200 间公司，其中 9%回答没有提价，5%有升价，86%没有回答问卷。调查结果是基于有回答问卷的 14%。

要注意样本偏差的证据，选错样本可能是无心，可能是有意。上文已提醒：样本是否足够的大，足以产生任何可靠的结论。

要同样小心处理报导的相关性：相关性是否够大，有重要的意义？是否有足够的案例赋予任何意义？一般读者不懂应用**显著性检验**[54]来确定样本是否足够。但许多报导一眼就能看出（可能要花点时间）是否有足够案例足以说服任何有理性的读者。

（三）什么不见了？

即使信息来源响当当，如没有明告有多少个案，已足以引起合理怀疑。同样的，如提到关连性但没有给出可靠性的计量（可能误差，标准误差），也足以引起合理怀疑。

提防平均值以及没有指明的各种平均值，要知道在很多情况，平均数和中位数会有很大差别。

很多数字没有意义，因为没有比较。例如「蒙古症研究发现 2800 个案例超过一半的母亲是 35 岁或以上」。除非知道妇女一般生儿育女的年龄，这说法没有特别意义。很少人知道妇女一般生儿育女的年龄。

另一例子：「卫生部最近公布的数据显示在过去雾霾天气的一周，死亡人数增加二百八十人。」死亡人数增加是否与雾霾有关？一般的死亡人数是多少？下一周的死亡人数会否减少？是否因为雾霾加速了某些人死亡？「死亡人数增加二百八十人」引人注意，但由于没有其他数字比较，意义不大。

如只给出百分比而没有原始数据，小心小心。很久之前，美国约翰霍金斯大学有一段有趣的报导：女大学生有 1/3 与教员共谐连理。惊人的百分比。原始数据说得清楚：许多年前，美国大学生只有极少数是女生；当年有三位女生，其中一人

---

[54] tests of significance

嫁给教师。

多年前，波士顿总商会的「优秀女性成就奖」宣称：十六位名列名人录的女士共有六十个学位和十八名子女。这些个人资料看来颇为扎实，但原本其中有两位奇人，她们共有三十个学位，而其中一位有子女十二人。

留意指数有许多疏漏：可能是基数。劳工组织指出在经济衰退后利润和生产指数上升快于工资指数。指数没错，但没有说明前者的基数较低，所以经济复苏时增加的百分比几乎必然是较高。

有时指数的缺失是没有说明导致变化的因素，有意或无意暗示是因为一些其他因素。今年二月的零售数字低于去年，但没有指出去年的春节是在二月，今年在一月。

过去几十年有关癌症死因的报告是误导的，因为有许多外在因素：以前对癌症所知不多，死因往往列为「死因不明」；现在有更多死因解剖，诊断更可靠，医疗统计数据较齐全；现代人更长寿，更多人活到容易患上癌症的年龄。如果只看总死亡人数而不是死亡率，不要忽视现在的人口比以前更多的事实。

（四）是否有改变主题？

留意原始数据和结论之间是否被转移，声东击西。

正如上文指出，更多呈报病例并不总是更多人染病。测验民意的投票并不一定反映正式投票的结果。杂志读者的兴趣调查不担保他们会从头到尾细读文章。

某年，美国加州中央谷地呈报脑炎病例大幅度增加，是去年的三倍。很多居民感到震惊，把子女暂送外地。但死亡数字没有很大改变；原来是州政府和联邦政府开始投入资源解决这个长期问题；因为他们的努力，发现许多以往被忽略的低程度病例。

大家可能留意到在某段时间，报章特多报导某类型的罪案或事件，感觉是无日无之，但过不了多久又沉寂下来。如仔细追寻，相关的官方数字没有增加。这只不过是有一两位记者当其时特别多这方面的报导，其他记者不得不追随其后。

英国公共工程部调查六千户有代表性的家庭，发表报告：「英国男士在夏天平均每周沐浴 2 次，冬天 1.7 次；女性是 2 次和 1.5 次」；引来报章头条报导英国男士每周沐浴次数多于女士。

这些数字要更令人信服，定要说明是平均数或中位数。然而，更严重的弱点是问题的主旨已经改变。调查真正发现的是「人们随口回答他们的洗澡次数，而这不是反映现实」。这是相当隐私的问题，受访者要顾全自己的面子（经常沐浴是良好的个人卫生习惯），对调查员给出的答案往往不是实际情况。

「离题」还有更多的品种变化。

《振兴农业》调查发现美国农场比五年前增加了五十万。这两个相应的数字其实不是计量同样的事情，因为调查局改变了农场的定义，新数据包括了旧定义不涵盖的三十万个农场。

人口普查发现奇怪的数据：例如三十五岁的人口不正常地多于三十四岁和三十六岁的人口。查究之下，发现数据是根据家人自报，他们倾向把岁数顺便调整为方便的五的倍数。要解决这问题的方法是要求呈报准确的出生日期。

中国某大区「人口」是 28 万，五年后升至 105 万。这幅度的增长当然有问题，深究之下原来两次调查是为了不同目的：第一次是税务普查，第二个为了饥荒救济。

美国也有一例。十年一度的人口普查发现 65 至 70 岁年龄组高于十年前的 55 至 60 年龄组。移民数字不能解释这差异。主要原因是颇大数量的受访者为了领取社会保障金而虚报年龄，也有可能是之前为了虚荣心而少报年龄。

美国参议员指责囚犯的住宿费用比市中心酒店还要昂贵，其实是混淆了囚犯的整体管理费用，这包括了监狱人员的薪俸。

各种事后孔明的废话是暗地改变主题的另一方式。

还有许多「我是第一」的形式。几乎任何事物都可以宣称自己是第一，只要不是太特别的什么。

当你考虑直接购买或分期付款，比较借钱成本会因为「改变主题」而难以比较。百分之六听起来像百分之六，但可能不是真的如此。向银行借贷 100 元，利率 6%，一年内每月清还利息约 3 元。但大多数汽车贷款标榜的「每百元利息六元」其利率实为双倍，不容易明白。

更糟糕的是美国的冷冻食品计划。粗心的买家被告知「6-10%」的数字。这听起

来是利息，事实并非如此。这是还款的数字，更糟糕的是这往往是以六个月计算，不是一年。100 元价格的食品，每月还款 12 元，等同真正利率 48%。难怪有这么多客户拖欠，食品计划要结束。

有时候会以语义来改变主题。《商业周刊》的报导：会计师决定「过剩」是讨厌的词语，提出企业资产负债表不再采用，改为「留存收益」或「固定资产增值」。

（五）是否有意义？

「是否有意义？」往往能够把基于未经证实假设的整个繁琐统计回归应有地位。Rudolf Flesch 提出文章可读性公式：简单和客观计算单词和句子的长度。以数字取代无法估量的论述，以算术取代判断，这是有吸引力的想法。至少雇用作家的人，如报纸出版商，甚至许多作家本身都应该注意。公式假设字词的长度决定可读性。这是否故意刁难，还有待证明。Robert A. Dufour 利用这公式评审一些文献，颇为得心应手，有助判断一篇文章、一本着作是否比较难读。

许多统计数字表面上已是虚假，只因为数字的魔力令人忘却了常识而蒙混过关。Leonard Engel 的多篇杂志文章列举了几个医疗案例。

一个例子是著名的泌尿科专家计算美国有八千万前列腺癌病例 - 足以涵盖易感年龄组的每位男性！另一例是神经科医生估计每十二名美国人有一人患有偏头痛；因为偏头痛占慢性头痛病例三分之一，这意味人人每一季度会患上失能性头痛。还有一个例子是经常提到的二十万宗多发性硬化症病，但死亡数据表明这种病例不会超过三至四万宗。

关于修改社会保障法一直饱受各种形式的声明；如未经仔细考证，这些声明各有各的道理。论点之一是既然预期寿命只有约 63 年，退休年龄订为 65 岁是虚假和欺诈行为，因为几乎每个人都在这之前死亡。

只要看看你认识的人就可以反驳这论点。基本谬误是这数字是指出生时的预期寿命，因此大约有一半婴儿可以预期活到 65 岁。顺便说一句，这数字来自 1939-41 年期间，已经过时但仍然使用。经过一代人后计算，目前的预测数字是 69.7 岁；这个新数字同样愚蠢，几乎每个人现在活到 65 岁。

多年前，一间大型家电公司的产品规划是基于出生率下降，长久以来已被认为是理所当然。规划要求重视小电器，适合公寓大小的冰箱。策划者之一突然回归常识：他放下图形和图表，转而留意自己和同事、朋友、邻居和旧同学，除了少数例外都有三、四个孩子或是计划大家庭。这重新启动没有成见的调查和制图 - 该

公司很快转向最有利可图的大户型。

赫然精确的数字往往违背人们的常识。纽约市报纸报导一项研究：与家人同住的在职妇女每周生活所需是 40.13 元。任何有常识的读者会意识到生活成本无法计算到最后一分钱。但是 40.13 元比「约 40 元」更动听，更是可怕的诱惑。

外推法[55]是有用的，特别是所谓预测趋势的占卜形式。看着这些数字和从中衍生的图表，必须记住：至今的趋势可能是事实，但未来趋势只不过是有些见识的猜测而矣。隐含的意思是「一切因素不变」和「目前的趋势继续」，但世事偏偏不会保持不变，否则人生会很无聊。

不受控外推法的废话，电视趋势是例子。在最初五年，美国家庭的电视机数量以百倍增加。依此趋势推论，再过五年会有几千万部，大概每家有四十部。

1948 年美国总统选战预测是统计史的大笑话。选举前的各项民意调查大多预测共和党候选人 Tom Dewey 获胜。结果是民主党杜鲁门得票 49%胜出。盖洛普选举预测被称为「人类历史上最公开的统计误差」。

专家分析民调出现偏差的原因，结论有三：调查抽样偏离了代表性、民调提早一星期结束，没能反映最后时刻的民意变化，以及政治偏见妨害了编辑的客观立场。当年报社老板多为共和党人，报纸挺共和党的当然较多。[56]

相对于一些未来人口预测，这已是准确的典范。近至 1938，总统的专家委员会深信美国人口永远不会达到 1.4 亿；十二年后这数字已是 1.52 亿。这些可怕的低估源于假设趋势将继续没有变化。

1874 年，马克・吐温总结了外推法的废话：

> 在一百七十六年间，密西西比河下游缩短了 242 英里，即是每年平均缩短 1 1/3 英里。依此推论，一百万年前的密西西比河下游足足有一百万英里长，像钓鱼杆伸出了墨西哥湾，也可以推论七百四十二年后，密西西比河下游将只有 1 3/4 英里。科学真有趣。只需投入少许事实就可以得出这样的回报。

[55] Extrapolation

[56] 改写自 http://hk.crntt.com/crn-webapp/mag/docDetail.jsp?coluid=36&docid=102284142&page=4

附录 香港大学民意调查的争论

（自学书院注：在翻译这本小书期间，香港正好有一场有关民意调查的笔战，也正好印证民调和统计的重要意义和容易陷阱〔正反双方皆如是〕。事缘香港特首[57]不是全民选举产生，无从得知究竟有多少选民属意他领导香港，于是定期民意调查是各方关注的寒暑表。香港大学民意研究计划和香港中文大学亚太研究所的定期民调最为各方关注。现任香港特首梁振英自 2012 年 7 月就任以来，民望一直在所谓合格线(50)徘徊。为此，行政会议[58]议员张志刚向香港大学民意研究计划发炮，引来一场不大不少的笔战。奇怪的是亚太研究所的民调结论也是差不多的「不合格」，但梁粉〔梁振英粉丝〕没有为此着墨。辑录这几篇文章颇多香港文体用语，请享用。）

港大民研发放特首及问责司局长民望数字

2014 年 3 月 11 日 〔香港大学民意研究计划〕新闻公报

特别宣布

在促进学术研究和理性讨论的基础上，香港大学民意研究计划（民研计划）今日在发放各项民望数字之余，更加把关键原始数据上载到《香港大学民意网站》，包括特首评分、被访者性别、年龄组别、以及加权指数。这种透明度，已经超过一般学术与专业要求，希望社会人士珍惜。学者专家使用及引用有关数据时，请按照学术惯例列明出处。

- 下载原始数据 ：2014 年 3 月 11 日公布之特首评分

公报简要[59]

民研计划在 2014 年 3 月 3 至 6 日期间，透过真实访员以随机抽样方式，成功以电话访问 1,017 名香港市民。调查显示，特首梁振英的最新支持度评分为 47.5 分，支持率为 25%，反对率为 56%，民望净值为负 31 个百分比，跟两星期前变化不大。…根据民研计划的标准，梁振英属于「表现失败」。 在 95%置信水平下，各项百分比的最高抽样误差为+/-4 个百分比，评分及支持率净值误差另计，调查的响应率为 66%。

注意事项：

[1]《香港大学民意网站》的网址为 http://hkupop.hku.hk，传媒可到网站参阅调查细节。

[2] 调查之样本为 1,017 个成功个案，并非 1,017 乘以响应率 65.9%，过去有不少传媒在报导上犯了上述错误。

---

[57] 香港特别行政区行政长官（又称特区首长、俗称特首；英语：Chief Executive）

[58] Executive Council，即是特首「内阁」。

[59] 这项定期的民意调查涵盖香港特区行政长官（特首）和主要官员的民望。为方便阅读，附录略去有关主要官员部份。

[3] 95%置信水平，是指倘若以不同随机样本重复进行有关调查 100 次，则 95 次的结果会在正负误差之内。传媒引用本调查的评分数字时，可以注明「在 95%置信水平下，各项评分误差不超过+/-1.8，百分比误差不超过+/-4%，净值误差不超过+/-6%」。由于民研计划在 2014 年引入「反复多重加权法」处理数据，交接期间，各项数字变化的差异是否超过抽样误差，是基于同类加权方法处理后的结果计算。换言之，2014 年第一次所得数据是否与上次调查存在显著差异，是基于两组数据同样经过反复多重加权后作出的比较，而非单从公布数字表面运算得来。

[4] 因为调查存在的抽样误差及处理数据的舍入过程，数字不能过份精确，合计数字亦未必完全准确。因此，传媒在引用有关调查的百分比数字时，应避免使用小数点，在引用评分数字时，则可以使用一个小数点。

[5] 调查数据并非透过音频互动系统取得，倘若调查机构以「计算机随机抽样电话访问」或类似文字来掩饰音频互动调查，是不专业的做法。

最新数据

民研计划今日发放特首梁振英及各问责官员的最新民望数字。2014 年起，民研计划把以往按照年龄及性别分布进行的简单加权方法，改良成为按照年龄、性别及教育程度（最高就读程度）分布的「反复多重加权」方法调整数据。今天公布的最新数据，是按照政府统计处提供之 2013 年底全港人口年龄及性别分布初步统计数字，以及 2011 年人口普查收集之教育程度（最高就读程度）分布统计数字，以「反复多重加权法」作出调整。现先列出最新调查的样本数据：

调查日期	总样本数	回应比率	最高百分比误差[6]
3-6/3/2014	1,017	65.9%	+/-3%

[6] 有关误差数字均以 95%置信水平及整体样本计算。95%置信水平，是指倘若以不同随机样本重复进行有关调查 100 次，则 95 次的结果会在正负误差之内。个别题目如果只涉及调查内若干次样本，百分比误差会相应增加。评分及支持率净值误差则会按照样本评分及支持率净值的分布情况另行推算。

由于不同题目涉及调查内不同次样本，误差会相应变化。下列参考数表笼统列出样本数目与最大抽样误差的关系，方便读者掌握有关变化：

样本数目（不论是总样本或次样本）	百分比误差[7]（以最高值计）	样本数目（不论是总样本或次样本）	百分比误差[7]（以最高值计）
1,300	+/- 2.8 %	1,350	+/- 2.7 %
1,200	+/- 2.9 %	1,250	+/- 2.8 %
1,100	+/- 3.0 %	1,150	+/- 3.0 %
1,000	+/- 3.2 %	1,050	+/- 3.1 %

900	+/- 3.3 %	950	+/- 3.2 %
800	+/- 3.5 %	850	+/- 3.4 %
700	+/- 3.8 %	750	+/- 3.7 %
600	+/- 4.1 %	650	+/- 3.9 %
500	+/- 4.5 %	550	+/- 4.3 %
400	+/- 5.0 %	450	+/- 4.7 %

[7] 以 95%置信水平计。

以下是特首梁振英的最新民望数字：

调查日期	2-6/1/14	15/1/14[8]	18-22/1/14	4-6/2/14	17-20/2/14	3-6/3/14	最新变化
样本基数	1,018	1,017	1,014	1,030	1,031	1,017	--
整体回应比率	66.5%	66.7%	67.6%	65.5%	67.8%	65.9%	--
最新结果	结果	结果	结果	结果	结果	结果及误差[9]	--
特首梁振英评分	45.6	48.9[10]	47.0[10]	47.9	46.4	47.5+/-1.5	+1.1
梁振英出任特首支持率	27%	29%	29%	25%[10]	23%	25+/-3%	+2%
梁振英出任特首反对率	58%	53%[10]	54%	56%	56%	56+/-3%	--
支持率净值	-31%	-24%[10]	-26%	-32%[10]	-33%	-31+/-5%	+2%

[8] 是次调查为施政报告实时调查，只问及特首评分及支持率。

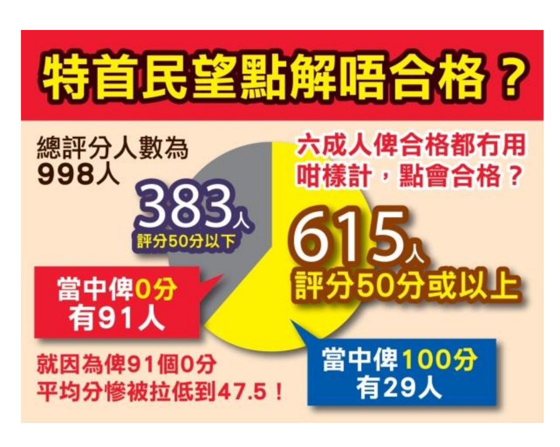
[9] 表中所有误差数字以 95%置信水平计算。95%置信水平，即是指倘若以不同随机样本重复进行有关调查 100 次，则 95 次的结果会在正负误差之内。传媒引用上述数字时，可以注明「在 95%置信水平下，评分误差不超过+/-1.5，百分比误差不超过+/-3%，支持率净值误差不超过+/-5%」；以前调查的误差数值请参阅网站。

[10] 该等变化在相同加权方法下超过在 95%置信水平的抽样误差，表示有关变化在统计学上表面成立。不过，数字变化在统计学上成立与否，并不等同有关变化的实际用途和意义。

【港人短评】解开特首民望「不合格」之谜

2014-03-14

港大民意研究计划的民调早阵子引起连串质疑，未知是否有见及此，今次港大再度公布特首评分时，民意网站已出现所谓的「原始数据」，虽然相关档案的格式要以特定软件才能打开，但内里所刊载的正正是评分分布数字。



民调应公正 做法须公平

依据港大最新的民调，以 100 分为满分，特首仅获 47.5 平均分，当然就被评为不合格了。然而，只要打开原始资料，就会发现 998 个评分者中，原来有多达 615 人、即逾 6 成人均给予特首 50 或以上的合格分数，其中更有 29 人给予 100 分；仅有 383 人给予 50 以下的评分。那么，为何特首的评分又会不合格呢？最大的问题在于有 91 人个受访者给予 0 分，就是这些极端评分，令特首的平均分大幅度拉低。

然而，这种意义甚为重要的评分分布，港大方面却未有主动公布，而只是藏在民意网站的暗处，若非主动寻找及装有特定软件，根本无法知晓！这种藏头露尾的安排，实在无法不令人怀疑民调背后的用意，即使不是存心误导，但这又是否一个公正持平的民调机构所应采用的发布方式呢？

收集及公布数据 必须高度透明

要知道的是，民调机构如何采用、公布、以至运用收集回来的数字，对最终的民调结果又或市民观感均起着决定性的影响。如此看来，香港确实有必要有更多独立的机构进行民调，并要高度透明地公布收集到的数据，以助市民大众通过比较获得真象。

张志刚[60]：六成二给特首打 50 分或以上说明什么！[61]

陈庄勤先生在 2 月 8 日于《明报》以〈沉默的螺旋〉为题撰文，对现时中大亚太所和港大民意研究计划所做的特首评分提出质疑。重点就是机构只公布平均分，但打分分数的分布却不清楚，只靠一个平均分，根本无法知道事情的真象。而本人上周撰文，指出单靠一个平均数，其实就是瞎子摸象。一般的研究，除了平均数之外，多会公布众数（最多人打的分数）、中位数，以及 50 分以上的比率。当时本人大胆推测，众数和中位数都是 50，给特首打 50 分或以上的应该超过一半。文章见报当日，港大民意研究计划也公布了最新的一次特首的评分，评分为 47.5，而港大也第一次以附录形式把所有评分的原始数据同时公布，这也是解决了陈庄勤和本人过去一直提出的质疑。因为附录必须要以 SPSS 软件才能打开，一般媒体都不具备这种统计分析的专用软件，所以没有引起广泛关注和报道。当我们打开这个原始数据档案时，马上真相大白。陈庄勤不用估，本人也不用猜。

港大首次公布所有原始数据

港大把给 0 分到 100 分的频率全部公开，可以说是非常公开透明。为方便表述解释，现把分数组合成 10 分一组，一共 10 组，评分分布见附图。

经运算之后，得出这样的结果。平均分是 47.5，众数是 50，中位数也是 50，给 50 分或以上的高达 61.8%。看完那些评分分布以及这 4 个重要指标，我们不需要再瞎子摸象，象的形状完全出现于我们眼前了！

平均分是 47.5，一般人的印象就是不合格！但如果看 50 分以上和以下的比例，在那 998 个给特首打了分数的人，有 28%的人打了 50 分，给 50 分以上的有 34%，那评 50 分以上的比率就是 62%，比 49 以及以下的 38%，多出一大截。当 62%香港市民给特首打 50 分或以上时，这是合格还是不合格？一些耸人听闻的讲法，例如民望破产之类，又从何说起。

把平均分拉到只有 47.5 分，最大的原因是大约有 9%的受访者打了 0 分。本人之前撰文也解释过，行政长官的施政有必然的两面性，无论政策多好，都会有一些人不满意。双辣招有八成人支持，但还有两成人反对，某程度是利益之争，持有多个投资物业的人就不支持，地产经纪也不支持，迷信绝对利伯维尔场的不支持。因为支持双辣招而支持特首的，可能给 60 分，但反对双辣招的就可能打 0 分。这种给行政首长的评分，就不能和读书考试相比拟，资质良好、读书用功的同学，

[60] 张志刚，香港行政会议（相等于内阁）成员，现任智库组织「一国两制研究中心」总裁。张志刚毕业于香港中文大学，分别获授学士及硕士学位，文章常见于本港各大传媒，着有《悲剧，悲香港》及《风雨声中》等书。

[61] 原文刊载于《明报》2014 年 3 月 18 日

可以科科取得优异成绩，甚至做 10A 状元。但行政首长推行政策，一定有得有失，结果也只会把平均分拉向中间。如果不看分布和其他指标，就只会以偏概全，甚至错下判断。

极端 10%主导舆情

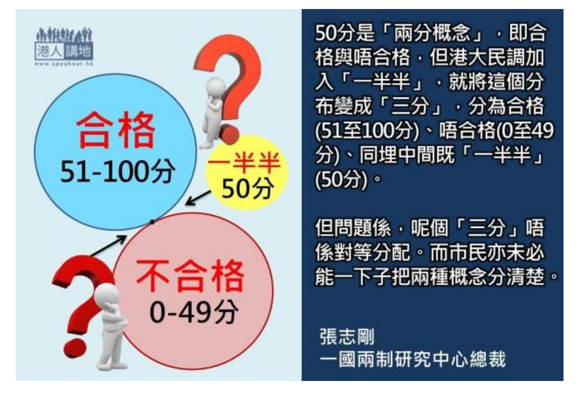
除了看那 50 分和以上占了 62%的重要数据，我们不妨再把那 10 组的分数逐一研究，0 分到 9 分的有 10.5%，这是最极端反对梁先生的一群。但 10 到 19 分的却只是 1.8%，20 到 29 分的也只有 3.9%。从分布来看，这不算是正常的分布，有点「恶之欲其死」的味道，到 30 和 40 分的两组，才回复正常，逐步回升到 8.9%和 13.1%。

给 50 分或以上的分布，就算是正常分布最多的是 50 到 59 分，占了 30.7%，愈高分数的比例愈低，逐步减少，没有出现 10 分和 20 分组别近于断层式的分布。而这一成给予 0 到 9 分的群组，相信也是最主动发声，最积极参与激烈行动的一群。当媒体的目光让这一成人吸引着，所谓舆情，便倾向了这最极端的 10%。50 分以上的组群，他们相对平和理性，政府施政，他们心中有数，但没有参与激进的意见表达活动，他们就成为了沉默的大多数。但当大学访问员来电时，他们就把自己的评价说出，但不幸的是，他们的评分又给那 9%给零分的人拉低冲淡，如果没有把所有得分公之于世的一日，这些沉默大多数的一群，永远没有见到「真象」的一日，也永远让那极端的 10%去主导舆情，和代表民情！

这种错误的代表，不仅是把民情扭曲，也形成了陈庄勤先生撰文中所提及的「白色恐怖的寒蝉效应」。支持梁先生的，支持特区政府的，都以为自己是少数，这令到他们变得沉默和冷漠，这也是反政府群体最希望见到的后果和现象。看完这堆港大公布的原始数字，真相大白于人前，支持梁先生的，支持特区政府的，不是少数！这说明过去一年半的政策走对头，证明特区政府官员的「勤力用心」，市民是看在眼里。

如果要正确的政策可以走下去，可以开花结果有成绩，不仅是需要市民打一个分数，更是要他们表达意见，更是要他们站出来！

张志刚：50 分应是「两分概念」



对于港大民意研究计划主任钟庭耀解释，民调中的 50 分代表「一半半」，即非合格，亦非不合格，一国两制研究中心总裁张志刚表示，以 0 到 100 分给分本来是一个「两分概念」，即合格与不合格，但港大民调加入了「一半半」，就将这个分布变成「三分」，即分为合格(51 至 100 分)、不合格(0 至 49 分)，以及中间的「一半半」(50 分)。但他质疑，问题是，此「三分」并非「对等分配」，而市民亦未必能一下子把两种概念分清楚。

练乙铮：打棍无效：网小子放倒「巨人」张志刚[62]

知识不等于力量，但如果缺乏知识，就可以很悲惨。无论在哪里，若统治阶级充斥不学无术之辈，社会大方向要出问题。这里说的知识，当然不是「公婆皆可有理」的看法认知，而是客观的学问。如果不仅是不学无术，还是别有心术的话，这个统治阶级无可药救。

卧虎藏龙

政改摊牌渐近，当权派集结力量围攻钟民调。先是政协委员、恒地副主席李家杰发飙，公开指摘钟氏经常在关键时刻发布对特府或北京不利的民调结果，操弄民意，为反对派开路。跟着，梁派网站《港人讲地》发表编辑室文章〈解开特首民望「不合格」之谜〉，指钟氏在最近的一个关于特首民望的民调里取巧运用数据说谎，把一个好端端成绩亮丽的特首说成多数人视为「不合格」。然后，梁派悍将、行会成员张志刚高调发言并在本周二《明报》撰文，引用上述网文核心内容，质问钟氏「六成二给特首打 50 分或以上说明什么？」【注 1】

结果，「六成二给特首打 50 分或以上」说明了《港人讲地》编辑室文章有「小小」搞错了基本统计方法，而「国师」张志刚懵然不知（？）并加小手脚发挥，结果闹大笑话。

最先指出《港人讲地》文章和张志刚说法有好几个严重初等错误的，是一篇又一篇的网上及新媒体文章，作者都懂统计，却是传统媒体里不见经传的业余评论者，可谓小孩大卫打死巨人高利亚，亦可谓：网络世界，卧虎藏龙。本文将这些材料整理，归纳所指出的谬误，并加若干己见，给大家参考。

首先指出，张志刚文章（下称「刚」文）的标题数字「62%」，与《港人讲地》编辑室文章（下称「讲」文）同源，是一个发水或抽水几近一倍的数字。「抽水」是指抽了民调响应者当中大批态度完全中立人士的水，把他们捆绑到梁特的支持者那边，便成功创制出上述那个发水标题数字。过程中，还擅自替民调加上一个不适当的概念，对所导致的矛盾和足令梁特尴尬的结论却讳莫如深。

张志刚的「62%」发水[63]近一倍

在港大钟氏民调里，特首「民望」数字的给定范围是 0-100，内含 101 个整数，

---

[62] 《信报》2014 年 3 月 20 日

[63] 发水：渗水发大

50 分居中。访问到的 998 个回应者当中，有 383 个给特首打的分数低于 50 分，280 个 50 分，335 个高于 50 分。钟民调事先给受访对象说明：「0 分」为「绝对唔支持」,「50 分」定义是 「一半半」, 100 分则为「绝对支持」。

因此，对统计者而言，必须严格尊重那 280 个打 50 分者的中立态度，既不能把他们摆到 383 个不支持者那边，亦不可将他们与 335 个梁特支持者放在一起；但是,「讲」文捆绑抽水好自便，把打 50 分或以上的访问对象加在一起（「一半半」+支持)，一算：（280+335）/998 = 62%，好亮丽！

然后张志刚就用这个数字说事，雄辩地问：这个数字「是合格还是不合格？」这就有趣了。因为这个算法如果说明特首民望是「严重地合格」，那么，我们同样可以把那 280 个态度中立打 50 分的受访者加到「不支持者」那边（「一半半」+唔支持)，算出（280+383）/998 = 66%。那不就表示梁特民望应该是「更严重地不合格」了么？

矛盾兼尴尬！正如一篇网文题目所说：「你玩统计，统计玩你」。**【注 2】**
任何公平的统计人，不会像「讲」文那样，抽那些响应「一半半」的态度中立人士的水，而只会用 335/998 = 34%这个数字，代表在原始数据里支持梁特的响应者比率。这个数字，固然比不上发水几近一倍的「62%」，与不支持梁特的回应者比率 383/998 = 38%相比，也差一截。如此，张志刚更应该雄辩地问问自己：34%这个数字,「是合格还是不合格？」

为何说事者可如此便给，大抽态度中立人士的水？因为中间做了几近无缝的概念转移。

政治态度中立 → 「合格」→ 「支持」

大家如果留意，当可察觉「讲」、「刚」二文其实歪曲了该次钟民调里的「50 分」的定义，把政治态度上的中立（「一半半」）巧妙地改成「合格」。然而这个民调里的 50 分，并非是一个「合格线」。

「合格」的标准人人不同。例如，笔者当年念的大学，合格线因教授而异；念津贴小学的时候，学校的合格分数是 60%；中学则是 40%，入读后，老父不满名校的标准反而那么低，笔者却认为好得很，因为可减轻功课做不好给老父指骂时的「杀伤力」。

然而，更重要的是，合格和支持不支持，其实没有必然关系——例如，某医学院专科生以 40.1%的分数合格毕业，你支持不支持这位仁兄当你的心脏手术医生？

「讲」、「刚」二文先将「50 分」擅自定义为「合格」（与民调对象回答问卷时的指定意义不同），然后再把这个他们引入的「合格」概念等同民调里的「支持」，这般偷换概念之后就可静鸡鸡进行上述捆绑抽水。如此，「刚」文就可大剌剌地说：「评 50 分以上的比率就是 62%，比 49（分）以及以下的 38%，多出一大截。」（注意：「50 分以上的比率是 62%」起码应该是「50 分或以上」罢？但连这个「或」字也省掉了。）如此逐步深入细致地做群众的思想摆布工作，不是第一次，大概也不会是最后一次。

事实上，港大民研计划已再三声明，「50 分」与「合格」完全无关，指的是态度上的中立。当然，可以有另外的民调专讲合格不合格，但这个梁特民望民调本身不适宜讲，硬要讲，就会出现上面的既矛盾也让梁特相当尴尬的结论。这个民调只研究特首民望的平均分数高低；得出一个平均分数之后，合格与否，读者可凭个人喜好各自解读。大概有些人，就算梁特民望拿个 1 分平均分，也会认为他是合格的；逻辑上，这没有问题，但如果滥用民调原始数据特别炮制一个「62%」来说事，就有问题。

剔除给 0 分的！保留给 100 分的！

所说何事呢？原来，「讲」、「刚」二文说，既有「62%」这个亮丽数字，而钟民调最后竟把梁特的平均民望评分算为 47.5，必是因为钟民调没有把打 0 分的那些「极端分子」——即统计学上说的「离群数据」（outliers）——剔除。于是，他们就可结论：钟民调不科学。这里有三个问题。

首先，如果要剔除给 0 分者，也应该剔除给 100 分者罢？但张志刚口中振振有辞的那个发水「62%」，却隐蔽地包含了 29 个「100 分」；这是「打茅波」。

其次，已经有专家算出，把响应分数最高和最低的 10%（含所有「0 分」和「100 分」）都剔除后，梁特民望平均值也好不了多少：48.1 分，救不了他；用张志刚的话说，依然「不合格」。如此，大动干戈为的显然不是两个平均分 48.1 - 47.5 = 0.6 分之差，因为「刚」文对此提都不提。那么，要剔除 91 个「0 分极端分子」，目的何在？不外起哄，令不谙统计学的人「觉得」钟民调无理。但请继续看无理的是谁。

第三，响应分数值既限在整数 0 与 100 之间，而 0 与 100 分在民调里都有清楚而具体定义，那么，根本就不应剔除响应值为 0 或 100 的那些数据，因为那些数据已经不能算是「离群数据」，而是民调设计者特别指明、更要知道的数据；理论上，0 分甚至可能是对象响应中的一个「众数」（mode）而意义尤其重要【注 3】。事实上，在该项民调里，给 0 分的 91 个响应，占 998 人的几乎 10%，相当于给

50 分的 280 个响应人数的三分之一；这许多响应者，怎可以看成都是该从统计数字里「枪毙」掉的呢？就看未加权的评分分布，我们也可以猜到，这个分布是双众数的（bimodal distribution），两个众数分别为 280 分和 0 分，因为的确有很多人对梁特极之不满；若取消了这部分人的数据，那就不是今天的香港了。统计学不应、也不允许那样搞出河蟹。

由此看出，不科学的不是钟民调，而正正是《港人讲地》编辑室和张志刚。
心术问题之外还有技术问题

「讲」、「刚」二文，还犯了一个技术性错误：「62%」这个数字，是拿了钟民调的原始数据做了小手脚就急不及待用来说事的结果，不知道人家有统计学的章法，就是对原始数据适当加权，之后才能用以作统计运算和分析。这里说的「加权」指什么？

大家知道，民调研究的对象人口总数太多，不能全部访问，只能抽样取板（sampling），但每一个随机样板中的个体特征分布如年龄、性别等，都不能准确反映总人口中的已知分布，此即所谓的「样板误差」；如果所调查的民意（如对梁特的态度）与年龄、性别等特征有关，样板便需加工，而统计学用的标准加工工序，是一个加权工序。笔者借用近日一篇网上流传很广、署名 SweetSourPork（「咕噜肉」）的《辅仁网》文章里的具体解释，稍作修改如下：

> 「如果今次电话访问，有 41.5%嘅受访者系男性，但系原来香港人口有 45.4%嘅人系男性，比受访者入面嘅男性多，咁我哋就要将样板入面嘅男性嘅比重加多啲，平衡番，等数据可以代表香港市民。」【注 4】

不做这个加权工序，样板误差可令民调的统计分析毫无意义。这是民调统计 ABC。「咕噜肉」于是用了钟民调的原始数据并作适当加权，重新再算一遍，证明钟民调算出的梁特评分平均数 47.5 没有错，错的是这里又犯了基本统计方法大漏的《港人讲地》和张志刚：那个已经包含抽水、概念僭建兼打茅波的「62%」，也是未经加权处理的（虽然因为前三个犯规动作太大太离谱，这第四个谬误相对而言已显得不那么重要）。

大家看看，一个饱含四个大错漏那么丰富的「数字」，尊贵的行会成员视为至宝，雄辩滔滔用来攻击对准钟民调。那不是很可笑吗？这种学养的人，放在本朝特府内外「智库」里打棍子很称职，安插在行会，则说到底有损其他大部分成员的面子和心理。港大民意研究计划成立于 1991 年，二十多年来，钟民调的学术功架已经十分娴熟，任凭当权派怎样抹黑，亦不能把他撼倒。最近这次围剿攻势，网民当中的专家见招拆招，已经代为瓦解。正如笔者早前提到，钟民调完全有资格

成为香港又一尊屹立不倒的图腾。

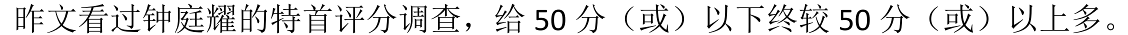
【注 1】李家杰言论见 http://zh.wikipedia.org/wiki/李家杰。《港人讲地》编辑室文见 http://speakout.hk/index.php/2013-11
-04-09-33-03/2013-12-21-08-43-26/1424-2014-03-14-10-38-16。张志刚文见 http://news.mingpao.com/20140318/msa.htm。

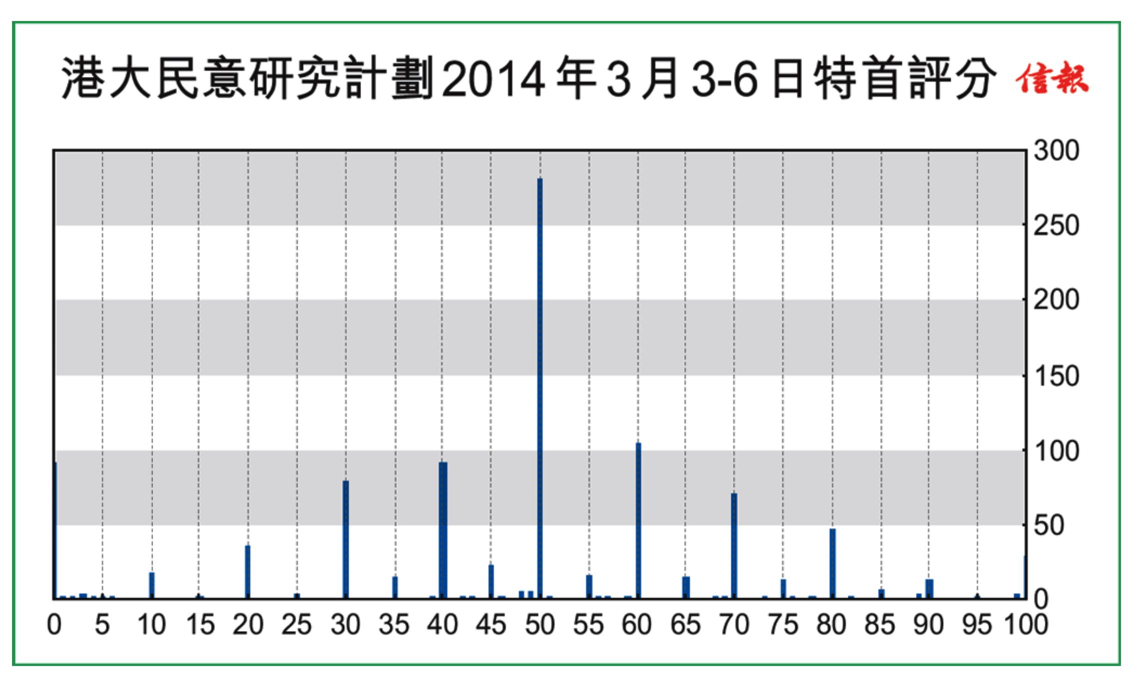
【注 2】见「陈电锯」的文章 http://www.chainsawriot.com/archives/9292；此文用另一统计加权方法（iterative sample bootstrapping），算出梁特的平均评分为 46.3，比钟民调的 47.5 稍低。

【注 3】关于离群数据，网文〈勿因虫废言〉 有很好的讨论：http://aloneinthefart.blogspot.co.nz/2014/03/blog-post_15.html；作者指出，一般而言，问卷响应若不设有效头尾限（例如 100 与 0）而是可以正负很大数以至无限的话，离群数据才有明显的潜在不良作用，应该剔除。文章分析头头是道，明显很在行；其上篇更值得看。

【注 4】「咕噜肉」文章〈港大民研特首评分系「被拉高」还是「拉低」？〉，用典型香港话写，解释统计过程深入浅出，见 http://www.vjmedia.com.hk/articles/2014/03/15/66322。不过，文章的加权评分分布图所表达的概念不对——应该是加权在人而不是加权在分，虽然算出的总平均分一样是对的。

罗耕：低水平的批评[64]


昨文看过钟庭耀的特首评分调查，给 50 分（或）以下终较 50 分（或）以上多。



说很多极端分子给 0 分吗？一样有不少给 100 分。难道全都要剔走吗？观乎分布，可能根本有些人想给超过 100 分，甚至有更多人想给负分，只是限于 0-100 无可奈何。如此 hit bound 的 tri-modal，用众数（mode）表达是无甚意思的，因这很可能是（-∞,+∞）的正态分布。假使调查的 50 分水岭改为 0 而两端不限，大概未必会见到这三峰现象。

在平均（mean）、中位（median）及众数三种中央趋势描述而言，若是量化数据，最可取是平均。当平均有机会被极端数字大幅拉高 / 低时，才用中位，譬如入息分布。然而，特首评分限于 0-100，无极端数字，故不宜用中位。只有 interior multi-modal 下，以众数表达多个中央趋势才有意思。至于张志刚指的 inter-quartile range，更不必了。

数据是否正态分布，其实可以 Jarque-Bera normality test 测试，详情可上维基看看。用原始数据不难算出，JB statistic 值达 386，显然呈正态分布。

批评钟庭耀的，看来要重新上基本统计课了。

---

64 《信报》2014 年 3 月 21 日

麦国华：民调科学与艺术[65]

回归十多年，特首民望时常被传媒打造成各具含义的大标题放在显眼位置，制造话题。如果说传媒为了吸引眼球而以文字渲染民调结果尚可理解的话，那么一间理应中立的学术机构若真的选择性公布某些调查数据，发布引导性结论，就实在令人为学术自由担心。

近日，港大民意研究计划遭揭发只公开有关特首支持度的「平均分」，而隐瞒原来有六成市民认为特首「及格」的事实，备受批评与质疑。然而，更让人为之瞠目的是民研计划负责人的反驳。他辩称「从来不会用 50 分等于及格去解释」，并称 50 分只是代表「中间意见」。

支持程度本就是一种相当感官化的心理状态，将其量化为具体数字，难免存在个人理解的因素。问卷设计者确可自行诠释不同数字含义，此问卷亦将 50 分定义为「一半半」，然该负责人过往曾解释「50 分以下等如不及格」，又何能自圆其说。加上某些自诩为香港良心的媒体也常以此为标准，疾呼特首民望不及格，大部分市民早被引导视 50 分为特首民望「及格」的界线。

面对质疑，该些媒体的反应更是令人心痛香港社会理智的流失。有媒体强调，揭出特首有 61%支持的是「梁粉」，暗示背后存在政治目的。一顶「梁粉」帽子就可否定一切证据事实。如此因人废言，和文革时期不问观点证据，单凭背景立场就批斗厮杀有何不同？

很多平日鼓吹公义平等的「道德卫士」们，攻击政府时高高举起，现在面对涉嫌违反公义的事情却又轻轻放低，彷佛事情只是桥下流水，其双重标准也应予诟病。倘若被指民调欠缺公允的是中央政策组或建制派的民研机构，恐怕早已尸横遍野了。只感叹，民调可以选择地公平，社会公义也可以选择地分配。

捍卫学术自由

捍卫学术自由，政府、市民、政党和学术界都有不可推卸的责任。民调的目的在于通过对大量样本的问卷调查来客观、精确地反映社会舆论或民意动向。民调结果会为政府所参考，从某种程度上可影响政府施政、市民心态及社会大环境。因此，市民有权利要求民研计划本着严谨的学术研究态度进行调查，全面客观地公布结果，让公道回归人心。遗憾的是，统计是一门科学，对统计数字的诠释，却是一门艺术。

---

[65] 《信报》2014 年 3 月 21 日

公说公有道，婆说婆有理？

「梁粉」批评如下：

依据港大最新的民调，以 100 分为满分，特首仅获 47.5 平均分，当然就被评为不合格了。然而，只要打开原始资料，就会发现 998 个评分者中，原来有多达 615 人、即逾 6 成人均给予特首 50 或以上的合格分数，其中更有 29 人给予 100 分；仅有 383 人给予 50 以下的评分。那么，为何特首的评分又会不合格呢？最大的问题在于有 91 人个受访者给予 0 分，就是这些极端评分，令特首的平均分大幅度拉低。

「主场新闻网站」及香港大学民意研究计划研究经理李伟健则反驳：
评论指有 91 个 0 分样本「拉低」平均分，没有提到 29 个 100 分样本同时会「拉高」平均分。港大民意计划研究经理李伟健向《主场新闻》解释，民望调查询问受访者给予官员 0 分至 100 分的评分，相信受访者诚实回答，无论样本是 0 分或是 100 分，都应纳入计算，除非是 101 分，在数值范围之外才会剔走。

李伟健强调，历来民望调查同样沿用这方法，公布按评分计算算术平均值（Arithmetic Mean)，「没有筛走特别低、特别高的评分。」

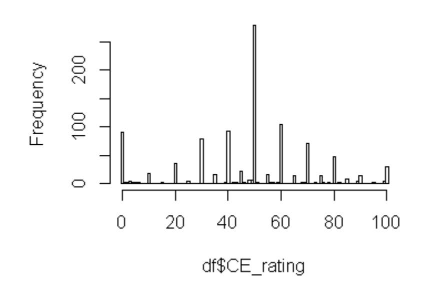
开门见山。我认为「梁粉」的批评有其道理，但其为己方所作辩解，一样有问题。另一边厢，「港大民研」的统计方法也有毛病。

Lies, damned lies, and 梁粉's statistics
统计数字不会说谎，它有的只是统计偏差。说谎的，是运用它的人。"Lies, damned lies, and statistics" 这句名言，就是用来讽刺那些蓄意运用统计数字来制造假像的人。前述「梁粉」的批评，正好拿来作「统计语言伪术」的最佳范例。

从「特首民望调查」所得到的 998 个有效评分，平均分为 47.4（「港大民研」公布数字为 47.5，略有不同，这是因为他们按受访者的统计特征作加权平均），低于 50，但实际上 998 个分数当中，**有 615 个为 50 分或以上**……至此，梁粉都没有说错。然而，他们没说的是：

998 个分数当中，**也有 663 个为 50 分或以下。**

感觉混淆吗？或者这样说吧，998 个分数当中，有 383 个低于 50 分，280 个等于 50 分，335 个高于 50 分。分数的分布如下：



从 0 到 100，共有一百零一个整数，而 50 正好居中。梁粉试图以「50 分**或**以上」这个标准来描绘一个梁振英有超过六成人支持的景象，可是据他们的逻辑，我们同样可以说，以「50 分**或**以下」这个标准来判断的话，有超过六成人（而且这个「超过六成」的人数比起梁粉的「超过六成」更多）反对梁振英！

我不明白一众梁粉何以如此介怀 47.5 这个只略低于 50 的数字。若是选举的话，两三个百分点也许是胜负关键，可是像印象分这种虽非玄学，却也「不算精密科学」的东西，47.5 和 50，实在没有分别。换了我是梁振英，看到如此数字，高兴还来不及呢。

离群值与平均数

梁粉指出，998 个分数当中，有 91 个是 0 分，这些极端评分拉低了整体的平均数。这是正确的。「主场」却反驳梁粉，说他们没提及样本当中亦有 29 个 100 分，会有拉高平均分的相反效果，也同样正确，亦再一次显示梁粉玩弄输打赢要的统计语言伪术。

然而，撇除梁粉的拙劣技俩不谈，若样本中可能有不少「离群值」(outliers) 的话，到底我们应该如何估计统计母体群的平均数？

港大民研的李伟健指「无论样本是 0 分或是 100 分，都应纳入计算」。就一般统计调查来说，这是过时的做法（但此处有一个 catch，要押后谈）。现代统计学认为「稳阵」(robust) 的做法，本网志之前的书评其实已经提过，就是利用截尾平均 (trimmed mean)，也就是先截去最高和最低的 5-10% 数据，然后才计算平均数。

可是我们几乎可以断言，在「特首民望调查」中，无论用普通的算术平均，抑或用截尾平均，都不会有大分别。原因是一般来说，离群值最有杀伤力的情况，是母体群数字本身为「无界」(unbounded) 的时候。是项调查当中，有效的评分本身有界（只可介乎零至一百），离群值的影响通常不会太坏，故此梁粉的批评，抓不到统计学的重点。

实际上，若截去今次样本当中，高低各一成的数据的话，得出来（未经加权）的截尾平均为 48.1，与样本平均数 47.4 相去不远。

尺度不同，分数如何换算？

这倒不是说「特首民望调查」无问题。印象中，港大民研所做的民意调查，大部份（例如立法会选举的选前调查和 exit polls）都很扎实。然而此项「特首民望调查」，却非常碍眼。我很想问钟庭耀一句：How on earth is this rating meaningful?

单单叫受访者为梁振英打个分数，已经很有问题。问卷只提过零分（「绝对唔支持」）、五十分（「一半半」）与一百分（「绝对支持」）的意义，中间的尺度 (scale)，人人却不同细分。你我各给六十分，意思未必相同。你的分数如何换算成我的，完全木宰羊。现时港大民研的做法，实际上假设了所有人的评分尺度均一。由此引起的模型风险 (model risk)，无法评估。举个例说，若你看到梁振英的「民望指数」比上月高，你可能以为他真的愈来愈受市民欢迎，但实情可能是他的民望无变，只是今个月的受访者的评分尺度较宽松，对无甚特别感觉的官员，也倾向打一个高分而已。

就算是奥运体操项目，评分有较多稍为客观的细项凭依（动作要求、难度、时限等等），仍不时惹人争议，各人对特首表现的评分尺度，又怎可能大致一样？

不知尺度，何论变化？

好了，就假设香港有一个平均的评分尺度吧。套用经济语言来说，就当人人都用一个一致「市场评分尺度」好了，但为何我们可以计算平均分？平均数并不一定是有意义的。一半人给零分，另一半给一百分，借用时下流行语来说，是社会撕裂的状况；所有人都打五十分，却更似人人认命。两种情况截然不同，平均分都是五十分，那么五十分究竟是甚么意思？

以上例子当然太极端，极端到与雷鼎鸣对坚尼系数的批评如出一辙。假若港大民研只是拿这个平均分来判断粗略民情的话，上一段的批评是不适用的。问题是，港大民研对待这个平均数时，彷佛其精密数值或它几个百分点的变化，有甚么微言大义似的。然而，即使香港有一个「市场评分尺度」，我们仍不知道这个尺度是甚么样子。同样是跌十分，从一百跌至九十分，是否跟六十跌至五十，或十跌至零同样大镬？木宰羊。五十分所代表的「一半半」，和「及格」是同样意思吗？木宰羊。不及格的话，甚么分数才算民怨沸腾，很想梁振英辞职？木宰羊。

不知背后的评分尺度的话，再精密的数字都是没用的。弄得好像很精密，反而令人误以为该数字很科学，其细微变化很有意义。

离群值真是离群值吗？
前面说过，以普通的算术平均来估计母体群平均数，乃过时做法。讽刺的是：

- 对「特首民望调查」来说，由于整把由零至一百分的量尺中，只有零、五十及一百有清晰意义，所以这三个分数，比其他分数可靠。
- 故此，**吊诡地，0 和 100 两个离群值，反而不应剔除。**
- 结果梁粉针对离群值的批评，意外地不适用。
- 若硬要计算平均数，普通的算术平均，此处亦反而比截尾平均更恰当。

然而这不表示港大民研的做法正确。正正因为他们采用了语意不明的尺度，才造成这许多奇怪状况。

结语一：less is more
如前述，港大民研的民意调查，一般都很扎实，但这项「特首民望调查」，用粤语来说的话，真系「畀位人插」。"Less is more" 这句说话听来陈套，但此处适用。奉劝 Robert Chung，还是干脆将问卷问题改成简简单单的「你想唔想梁振英继续执政」之类好了，不要再搞那些懒细致的评分吧。

结语二：废话去死，自由万岁
最后且谈文字，不谈统计。梁粉谓：

港大民意研究计划的民调早阵子引起连串质疑，未知是否有见及此，今次港大再度公布特首评分时，民意网站已出现所谓的「原始数据」，**虽然相关档案的格式要以特定软件才能打开**，但内里所刊载的正正是评分分布数字。
这不是废话吗？有甚么档案是任何软件都可以打开的呢？何况所谓「特定软件」和文件格式，也不过是统计佬惯用的 SPSS 与它的 sav 格式吧。不想付钞的朋友，可用免费的自由软件 R 打开有关档案。

相关网页

- The R project for statistical computing
- 2014 年 3 月 11 日 新闻公报；香港大学民意研究计划
- 下载原始数据（SPSS 的 sav 格式）：2014 年 3 月 11 日公布之特首评分
- 民情指数方法说明 (pdf)；香港大学民意研究计划

伸延阅读

- 电锯，你玩统计，统计玩你：「问题根本不在于 0 和 100 等等 outliers ，而是占人口比重较多的组群对梁振英评分较低。」

请钟庭耀回应 请关焯照澄清 / 文：张志刚

(明报) 2014 年 03 月 25 日

由前周钟庭耀公布了特首评分的原始数据之后，就引起广泛的分析和讨论，这其实是好事。学术机构的行为，理应面对公众批评，不要随便就以「抹黑」和「打压」视之。而关焯照先生等也写了一篇专文，提出不同意见，个人在此尝试把事情详细再分析一遍。关先生和其他有兴趣的人士可以详细阅读思考，往后可以再作交流或者当面讨论。

整件事件似是复杂，但如作有条理的梳理，其实不难掌握。关键是钟庭耀的特首评分，有没有合格的概念和应用。此关键一解，往后就是大路一条。

钟庭耀在 3 月 19 日接受《信报》访问，指出「50 分是中位数，不能演绎成正向或负向数字，从来不能说 50 分合格」。

钟庭耀的解释，涉及两个问题，一是这种评分，有没有合格嬲尸 X 格的概念。二是如果有，又应该几多分合格。

钟庭耀的评分，其实做了很长历史，太远的不说，就从回归谈起，也有 17 年。这 17 年来，媒体从来都以合格嬲尸 X 格的概念来报道特首评分，而且都以 50 分为合格。香港媒体事业发达，每次数字一出，电视、电台、报章都踊跃报道，这合格嬲尸 X 格词语，出现起码 100 次。钟庭耀每月起码做一次调查，1 年 12 次，加起来就过千次。17 年来，少说也报了一两万次。如果钟庭耀认为这个调查根本没有合格嬲尸 X 格的概念，那在过去 1 万多次的报道，钟庭耀为什么不挺身而出、拨乱反正？就在前周公布原始数据之后，得出评 50 分或以上有六成二人的结果，钟庭耀才急忙表态，认为没有合格不合格，又或者 50 分不能视为合格之说。

曾被引述 50 分为及格水平

香港的记者、编辑，多是有识之士，他们一个错不奇，个个都出错？他们视 50 分为合格，固然是凭自己的固有认知，而钟庭耀自己也有不可推卸的责任。因为他仙人指路，他本人就是如此演绎。本人的一位同事用了一个下午的时间，在慧科电子剪报搜寻过去 10 多年的相关报道，找到以下这些材料。请记着，这些报道是直接经访问引述或直述钟庭耀的分析，而不是媒体自己的报道。如果只计媒体报道，那是成千上万，不必在慧科电子剪报搜寻。

《苹果日报》2010 年 8 月 11 日：「民意研究计划总监钟庭耀分析，按曾荫权的民望表现而论，他的民望属『表现失败』。虽然他的评分有轻微上升，仍可以维持在略高于 50 分的及格水平。」

《头条日报》2010 年 7 月 28 日：「该研究计划总监钟庭耀表示，虽然曾荫权评分脱离肥佬行列。」（注：评分为 50.3 分）

《星岛日报》2004 年 10 月 13 日：「钟庭耀认为他（杨永强）的支持度保持稳定，比其历史低位 39.4 分高出很多，但仍未达到 50 分的及格水平。」

《星岛日报》2004 年 9 月 29 日：「钟庭耀分析，调查结果显示董建华的民望评分两年来首次重上 50 分水平。」

《信报》2003 年 9 月 10 日：「钟庭耀指出……孙明扬……杨永强……林瑞麟……马时亨全数低于 50 分的及格水平。」

《明报》2003 年 8 月 13 日：「钟庭耀分析：『……余下 12 个问责官员中只有 4 个不及 50 分，算是初步走出管治危机。』」

《明报》2003 年 1 月 29 日：「钟庭耀指出，特首评分自去年 8 月起已连续半年处于不及格水平……连续半年处于 50 分以下。」

另外慧科电子剪报显示 2003 年 9 月 24 日和 2004 年 4 月 14 日的《苹果日报》，在为特首和主要官员评分制表时，分别出现「注：评分以 50 分及格」（2003 年 9 月 24 日）、「注：评分由 0 至 100 分，50 分及格」（2004 年 4 月 14 日）等字样，并且都写明「数据源：港大民意网站」。

钟庭耀 1997 年 7 月出版的《民意快讯》第 11 期，在总结港督彭定康的支持度评分时表示：「整体而言，彭定康所得的分数一直能够维持在 50 分的合格分数以上，反映彭定康在市民心目中的形象尚算不俗。」据港大民意网站介绍，无论是对回归前的港督，还是回归后的特首，支持度评分的提问方式是一样的。

任何稍懂中文的人，也可以从上述的引述，清楚理解，这套评分方法是： 0 至 100 分，50 分为合格。讲了千次万次，钟庭耀自己也是如是说。今日被翻出有六成二的人给了梁振英先生合格的分数，就走出来完全推翻过去 17 年的定义，作为香港大学的民意调查机构，钟庭耀是不是要正式响应？

看完以上的引述，相信已经可以解答了关焯照先生的问题，但为求详细交代，以

下再作进一步的分析。关先生等 3 人是懂得统计的人士，请 3 位首先思考并回答一个问题：钟庭耀的评分，是归类为定序（Ordinal）还是定距（Interval）的问题？所谓定序，通常是 3 项式选择，响应者独立挑选，只能每选项独立计算频率，选项之间也不存在空间可供选择。中大在 2012 年初对候任行政长官支持度作调查，就提供了 3 个选项：不支持、普通嬰@半半、支持，这 3 个就是回应者可选的答案。在计算机运算时是用代码，但运算后出来的答案结果仍然是不支持、普通嬰@半半、支持。如果是定序（Ordinal）的问题，我当然不能把一半半的归类为支持，这是不能接受的错误，这种方法也同时不能相互运算，所以不会有平均分这结果。

看钟庭耀问卷的问题，是典型的定距（Interval）的问题。0 至 100 是连续，不是独立方块。数字可以相互运算，所以有平均分的出现。如果关先生用 SPSS 查看钟庭耀的原始数据，可以发现答案只是出现 0 至 100 分，从来没有不支持、一半半、支持的字样。这 3 组字不是答案，而只是用来向受访者解释 0 至 100 分的方向和意义。这个所谓一半半，在统计学上，和上述中大那个一半半，两者完全不同意义。在定序（Ordinal）里，一半半是独立成章，本身就是答案。但在定距（Interval）中，50 分就是 50 分。而一般人对 50 分是合格分的印象已是根深柢固，早有定论。再加上媒体的报道，以及钟庭耀自己也不断解读 50 分为合格分，所以本人以 50 分为合格分起点，向上计算得出 62%之数，又有何问题？如果真的要重回一半半的本来意义，那就只能用回中大那个问题，一半半独立成章。但如果用 3 选项而不打分数，又无法制造「民望肥佬」的形象！

「平分春色」欠基础
此外，关先生也提出把给 50 分的频数一分为二，一半拨入支持，一半拨入反对，平分春色。

关先生这种做法，是完全混乱了取态上的一半半，和人数上的一半半。真的要知道给一半半的响应者的最后取态，就只能在访问中再追问一条问题：「如果没有一半半可选，那是会投入支持，还是投入反对？」另有一可能就是弃权不选。转投的比例，根本无从得知，可能是八对二，也可能是三对七，我们凭什么基础去假设五成对五成？推论可以接受，但总要有一些基础，例如参考其他两分法民调的结果，而不可以随意一分为二，这点希望关先生可以澄清。归根究柢，我们必须清楚评分本身就有合格嬰尸 X 格的概念。而且一定有一个划分点（cut-off point），而没有中间形态。合格就合格，不合格就不合格，刚刚合格的下一个分数就是不合格，就是这么简单。

后记：默书拿 50 分的儿子问妈妈：「妈妈，我合格定唔合格？如果 50 分不算是合格，由 51 分才算，那 50 分又算什么？又是合格，又是不合格？不能算是合格，

又不能算是不合格？」几经折腾，妈妈最后无奈叫儿子：「你去问钟 sir！」这时，妹妹跑过来告诉妈妈：「我默书也是 50 分，合格还是不合格？」妈妈喜形于色回答：「你哋一个合格，一个不合格。」（文章仅代表个人立场）

〈潮池 Blog〉画出肠民调之子矛子盾计

不胜其烦，有关特首民望调查的争论，无奈继续。

港大民意研究计划负责人钟庭耀澄清，50 分在特首评分中，在问卷问题上，定义为「一半半」，统计学上属「中间数」，不应视 50 分为「合格」或「不合格」(详见〈画出肠民调之一池浑水〉)，张志刚在《明报》一文〈请钟庭耀回应，请关焯照澄清〉，试图以子之矛，攻子之盾，谓多年来，报章最少九次引述钟庭耀形容「50 分为及格水平」，以证钟庭耀自打嘴巴。

实情如何呢？

因为要准备是日香港电台《自由风自由 phone》节目，笔者用「慧科搜索」，复核了该文九个试图指控钟庭耀自打嘴巴的「例证」，功课已做，乐意公诸同好。

文字的确存在，不过……

(如果大家觉得好烦，请跳过以下二十三段，从尾六段开始看总结就可以了。)

(以下九「例证」引自张的文章)

「例证一」:《苹果日报》2010 年 8 月 11 日：「民意研究计划总监钟庭耀分析，按曾荫权 的民望表现而论，他的民望属『表现失败』。虽然他的评分有轻微上升，仍可以维持在略高于 50 分的及格水平。」

评：当天共有六份报章有引述钟庭耀分析，只有《苹果日报》提到他说「仍可以维持在略高于 50 分的及格水平」。1. 有可能是记者引述不精准，也有可能是钟庭耀这样说。2. 按前文后理，「仍可以维持在略高于 50 分的及格水平」有歧义，可诠释为「50 分」是及格水平或「略高于 50 分」是及格水平。

「例证二」:《头条日报》2010 年 7 月 28 日：「该研究计划总监钟庭耀表示，虽然曾荫权评分脱离肥佬行列。」(注：评分为 50.3 分)

评：「脱离肥佬行列」，如何诠释为「50 分为及格水平」？

「例证三」：《星岛日报》2004 年 10 月 13 日：「钟庭耀认为他（杨永强）的

支持度保持稳定，比其历史低位39.4 分高出很多，但仍未达到50 分的及格水平。」

评：当天共有八份报章有引述钟庭耀分析，只有《星岛日报》引述钟庭耀就样说。有可能是记者引述不精准，也有可能是钟确实这样说过，难证实。中文大学的同类调查以五十分为「及格」，有可能令少部分记者也诠释港大民研调查的五十分为「及格」。

「例证四」：《星岛日报》2004 年9 月29 日：「钟庭耀分析，调查结果显示董建华的民望评分两年来首次重上50 分水平。」

评：当天共有十份报章有引述钟庭耀分析，都有类似字眼，但「重上 50 分水平」，不可能解读为「50 分为及格」的意思。正如评分「重上 60 分水平」，不可能解读为「60 分为及格」。

「例证五」：《信报》2003 年9 月10 日：「钟庭耀指出……孙明扬……杨永强……林瑞麟……马时亨 全数低于50 分的及格水平。」

评：上段引述有很多省略号，原文是这样的：

「钟庭耀指出，市民对财政司司长唐英年及保安局局长李少光的评价相当不俗，可见人事更替似乎可以为政府带来一点好处。不过，接替唐英年出任工商及科技局局长的曾俊华由于市民认知率不足三成而不获排名。

房屋及规划地政局局长孙明扬、卫生福利及食物局局长杨永强、政制事务局局长林瑞麟和财经事务局局长马时亨全数低于五十分的及格水平，以林瑞麟及马时亨最低分，分别有四十三分及四十二点九分。」

正常新闻写法，很明显最后一段并非引述钟庭耀，「五十分的及格水平」属记者自己的诠释。「例证五」的省略号省得太多了。把两段文字砌埋一齐，改变了意思，这就叫「断章取义」。

「例证六」：《明报》2003 年8 月13 日：「钟庭耀分析：『……余下12 个问责官员中只有4 个不及50 分，算是初步走出管治危机。』」

评：按当时诠释的前文后理，钟庭耀一直以45 分为「信任危机线」，故有此说。而「不及 50 分」之讲法，亦不能视「50 分为及格水平」。

「例证七」：《明报》2003 年1 月29 日：「钟庭耀指出，特首评分自去年8 月起

已连续半年处于不及格水平……连续半年处于 50 分以下。」

评：这是较离谱的一个引述，翻查原文，上述引文的省略号，省了三大段。原文第一段是「「港大民意网站」发现，特首董建华的民望，由 1 月中的 47.3 分跌至 1 月底的 45.2 分，下滑 2.1 分(若综合其他数据，1 月平均分为 46.3 分，见图)，再见历史新低。民意研究计划主任钟庭耀指出，特首评分自去年 8 月起已连续半年处于不及格水平，反映政府有管治危机。」

然后隔了三段，才是「民意研究计划主任钟庭耀认为，特首民望自去年 8 月起，连续半年处于 50 分以下，并屡创新低，情况前所未有。」

而且，按钟的说法，50 分以下，属不及格水平 (50 分为一半半，50 分以上为及格)，此文与钟的一贯讲法无矛盾。如此拼凑证据，制造错觉，唉。

「例证八」:另外慧科电子剪报显示 2003 年 9 月 24 日和 2004 年 4 月 14 日的《苹果日报》，在为特首和主要官员评分制表时，分别出现「注：评分以 50 分及格」（2003 年 9 月 24 日）、「注：评分由 0 至 100 分，50 分及格」（2004 年 4 月 14 日）等字样，并且都写明「数据源：港大民意网站」。

评：不能排除「评分以 50 分及格」为记者的诠释，在港大民意网站中，找不到「评分以 50 分及格」的字眼。找到的请告诉我。

「例证九」：钟庭耀 1997 年 7 月出版的《民意快讯》第 11 期，在总结港督彭定康的支持度评分时表示:「整体而言，彭定康所得的分数一直能够维持在 50 分的合格分数以上，反映彭定康在市民心目中的形象尚算不俗。」据港大民意网站介绍，无论是对回归前的港督，还是回归后的特首，支持度评分的提问方式是一样的。

评：翻查港大民研出版的当期《民意快讯 》，确实清楚写到 50 分为及格分数的说法。这是九个「例证」中，唯一一个清晰见到有「50 分为及格水平」的字眼。钟庭耀如果要奉陪辩论下去的话，这点需要解释。笔者意见，港大民研网站如大海一样的历史资料，只有一两个矛盾位，「算系咁」。

长篇大论，真的唔好意思。总结：九个「例证」，五个为曲解、误解或过分跳跃阅读的错解，三个有可能是记者自己的诠释，只有一处 1997 年的说法出现矛盾。

张志刚与建制派的批评，一直针对港大民研计划，其实中大也一直有同类型调查，为何不批判中大呢？他们要求要公开调查原始数据，港大民研自负盈亏，数据属

学术资产，是日最新发展，港大民研发声明，公开全部有关梁振英民望的原始数据，真的慷慨。其实，中央政策组也用公帑资助不少学者做研究，他们的研究成果，枉论公开原始数据，研究报告也只能于网上查阅到摘要。既有此「公开原始数据」的要求，是否公帑资助的研究，也应公开原始数据？

统计数据，应用 interval 还是 ordinal，各有优劣，50 分应如何定义与诠释，本应属于学术讨论范畴，难分对错，而且任何方式的诠释，也只差两三分，为何左报与建制舆论对一个学者频密施袭了？大家何时对学术咁有兴趣了？

事件风眼中的主角钟庭耀，一直甚少正面响应各种批评，他最近在港台《传媒透视》有一篇文章〈从国王的新衣的说起〉，详细说了「国王的新衣」故事，文末有这样两段：

「国王没有雅量，谗臣乘机取巧。先把小孩打成造反派，再把科学变歪理。然后口诛笔伐，肆意攻击，制造白色恐怖，以为可以解决问题。谁知道，真理不会被改变。掩耳盗铃，只会弄巧反拙。

面对来势汹汹的攻击，笔者并不急于响应。有助学术研究和公民社会发展的理性讨论，笔者当然积极参与。对于那些不怀好意、借故诋毁的谩骂，就由它们在历史洪流中消失好了。真理不在口舌之间，只要把事实纪录下来，谁是谁非，历史自有分晓。」

民调　真相此中寻　[关焯照、周文林、雷照盛]

苹果日报　2014 年 3 月 26 日

港大民意研究计划（下称「港大民研」）的特首民调争议越演越烈。网站「港人讲地」和行会成员张志刚在这几天仍在电子传媒和报章发表批评，认为港大民研以评分计算民望的做法有问题。同时，将 50 分厘定为「一半半」可被一般市民视为合格分数，此外，将被访者的评分划分为「0 至 49 分」、「50」及「51 至 100 分」的概念，可能令问题含糊化。

首先，笔者写这篇文章的目的是，（1）澄清一下做民调分析需要注意的地方，（2）希望避免民调结果的解读产生误解。

港人讲地及张志刚猛烈批评的港大民研的民调问题是特首的支持度评分，其的内容是：「而家想请你用 0-100 分评价你对特首梁振英的支持度，0 分代表绝对不支持，100 分代表绝对支持，50 分代表一半半，你会畀几多分梁振英呢？」

港大民研是采用统计学上常用的等距量表（Interval Scale）的方法去量度特首的支持度（由最低的 0 分至最高的 100 分）。这种做法的好处是从得分上了解到市民支持特首的「程度」（附图）。大家可以细想，有两位被访者给予的分数是 51 分和 90 分，显然，评 90 分的被访者的支持度远较评 51 分的被访者为高，但如果采用港人讲地和张志刚的提议方法去分组，以 50 分为中间点分界，然后将 0-49 分和 50-100 分别厘定为「不合格」和「合格」，读者便不能看到这两个评分的差异了。

港人讲地和张志刚的做法是将 0 至 100 分的范围变换为两个不同组别，「合格」与「不合格」。如果用统计学的说法，他们是用一个顺序量表（Ordinal Scale）去将数据分类——即是变为分类数据。如果用以上例子，51 分和 90 分是纳入为同一组别（合格），但问题是 51 分和 90 分是分代表不同程度的支持，但在归纳组别过程（Aggregation）中，这种支持程度的差距便会被剔除，对研究者来说，这可视为流失了重要资料，最终令研究质量被拉低。

一个相关的难题是一旦采用港人讲地和张志刚所提出的二元答案（合格和不合格）作为分析，在这情况下，问题的字眼和答案是需要修改。例如，问题可写为：「你支不支持特首梁振英？」而答案分别是「支持」、「不支持」和「无意见」。一旦港大民研的问题重新改写为港人讲地和张志刚的问题格式，得出来结果（例如支持度的百分比）是极可能有差距，因为问题的本质和问法已不同，至于差距

在统计学上是否有明显分别，这便要用适当的统计方法去验证了。

最后，另一个争论点是 50 分是否一个合格分。单以民调的问题措辞，笔者看不到港大民研有任何表示 50 分是一个合格分数。至于「一半半」，是一个中性词汇，可解读为「中间点」、「一般」、「普普通通」等。然而港人讲地和张志刚坚持认为 50 分是一般人理解为合格分数，这只是他个人意见，正确与否，学界自有公论。

现在整个港大民研的民调争议只是各说各话，犹如鸡同鸭讲。但笔者要指出，做学术研究是需要保持严谨态度，无论从民调内容、样本的收集方法和统计分析均要达到起码的学术水平，这才能令人信服。

关焯照　经济学家、冠域商业及经济研究中心主任
周文林　经济学家、冠域商业及经济研究中心高级研究员
雷照盛　统计学家、港大统计及精算学系讲师、冠域商业及经济研究中心研究员

卢先亚：特首的妈（一）

2014-3-28

前几天看到了张志刚先生为了护主，在他报再次向钟庭耀博士及挺身而出的关焯照博士，就民调一事「叫阵」，且在文中引述好些统计学的专业用语，例如甚么等距 （interval）、有序（Ordinal）数据等等， 明显就是要吓唬外行人。我自问不学无术，未敢轻言反驳，所以特地请教我的一位学弟，现该说是一位学者。他与统计结缘廿多年，持有统计学博士学位，年少时甚至当过访问员，及后任教统计课程，并主理多个大型统计调查及参与民调工作，现仍在这领域继续研究，可知其醉心程度。

当我致电并道明来意，他努力尝试透过电话解说，我越听越唔知佢嗡乜，咁话晒都系学究嘛，当他亦然发觉话筒另端的「接收」 有问题，他说不如发个电邮以资说明，我自是求之不得。虽然我还得再三恳请他要写得浅白入屋一些，而他亦同时叮嘱我千祈「唔好开名」。我明白学院中人大都不爱抛头露面，惟更清楚的是，若然无端拖他下水， 只怕钟庭耀之外，又多一位统计专才遭受打压，我又于心何忍。不过，跟手收到其洋洋数千字的鸿文更知， 其实佢根本就系想直斥痛骂张志刚！我又怎不玉成美事。惜原文太长，节录之余，还要分日列出。以下是学弟的话，而括号内乃我后加：

张志刚先生，在此响应你在报刊所写。特首也并不是我的儿子，我更不愿作特首的妈！（谁又想天天捱骂呢！）一区之首亦不是小朋友默书考试！我不知道阁下对儿女要求如何，但大部分港妈亦不会接受仔女只拿 50 分，何况是特首要职！比方说，在职场上，谁会接受在工作上只有 50 分的下属？怕早给炒掉了！（这点我可左证）大部分有志气有理想的人（与张先生无关），亦不会甘心跟随能力只有 50 分的上司工作，没前途的吧！所以请不要在 50 分上沾沾自喜，况且我们的特首在最新的港大民调中只得 47.5 分呢！

在张先生文中，论定港大民调问卷中的所谓支持程度是属于等距(interval) 数据，原因是原始数据(raw data) 只记录了 0 至 100 分，当中并没有支持、一半半及不支持的字样。这种论证确实粗疏！专业统计人员都知道，原始数据不能单独使用，一定要参照编码手册(coding manual) 或问卷设计。举例，问卷可能会包含一些有关出生地、职业、行业等问题，一般会用数字代码记录（例如 1 代表香港、2 内地及 3 其他地方），一来比较方便，亦同时大大减少电子档案存量。如果不参照编码手册 （coding manual）或问卷设计，原始数据就出现不能解读，甚或误读的情况。而张先生的论据只是简单对号入座的误读罢了。

参考港大民调问卷，该问题是：「而家想请你用 0 至 100 分评价你对特首梁振英既支持程度，0 分代表绝对唔支持，100 分代表绝对支持，50 分代表一半半，你会俾几多分特首梁振英呢？」自 90 年代起，港大民调一向是使用 CATI 系统（学弟列出全写， 我从略），即是使用计算机抽选电话，自动拨号至接通，访问员会准确依据计算机所示读出问题再把受访者答案输入计算机，整个过程亦有主管在旁监听以确保数据质素。所以每个受访者亦会清楚明白 50 分代表一半半， 而不是代表合格，这是无可争议的。

卢先亚：特首的妈（二）

2014-3-31

在讨论甚么是合格之前，首先要了解甚么是支持程度。支持程度和考试测验最大的分别是后者大多数有明确的评分标准，例如答对一题有 10 分，而合格标准则是老师或教授们的专业判断。学术程度越高，合格标准就越严格，例如医生、工程师等专业考试要求就很高，人命关天嗅！所以考试分数大多是定义明确的集合（well-defined set） 。但在社会研究或行为科学等领域中，很多时要处理一些含糊不清、定义不明确的变量（variable） ，数学上称为模糊集合（Fuzzy set ），例如快乐、情绪、生活满足（life satisfaction）、工作动力 （work motivation）等等。一些社会学家、心理学家、计量心理学者 （psychometrician） 、教育学者就会以李克特量表（Likert Scale，下简称量表） 为这些模糊概念作简单的量化描述，即是问卷常用的 5 级设计 ：

1. 非常同意
2. 同意
3. 既不是同意亦不是不同意（或作中立）
4. 不同意
5. 非常不同意

有些研究员会再把量表扩展为 7 级或更高级别，而港大民调只是把量表以 0 至 100 分表示，而 50 分则为 101 级量表的「一半半」！ 对照 5 级量表其实分别不大，只是支持及不支持两方面被划分得更仔细。值得注意的是，量表并非等距，即是 （4 不同意）并不是（2 同意）的两倍，但一定对称（symmetric）。同理，港大民调中所谓的支持程度，50 分亦不是 25 分的 2 倍，而用量表所计算出来的平均数亦只是一种中间趋势的描述，这亦是对称设计的结果。

那么怎样才叫合格？钟博士讲得很清楚， 在港大民调设计之中并没有考虑这问

题！至于怎样去订立合格线，我建议可在港大民调中加入问题，例如问：你觉得作为一个特首，社会大众对其支持程度（0 至 100 分） 应该（i）要达到几多分以上才可以叫做合格（即 Pass）呢？（ii）要达到几多分以上才可以叫做良（即 Pass with Credit）？ （iii）要达到几多分以上才可以叫做优（即 Pass with Distinction）？ 另外， 亦可找来政治学及公共行政学的学者（经济学者，尤其姓雷的，大可不必）们，为特首这职位定一些标准。当中并不一定只采用社会大众的支持程度作唯一条件，同时可加入其他可测计量，例如 GDP 增长、坚尼系数、犯罪率、环保指标、新闻及言论自由指标等等。

我只想强调，特首是重要之职，合格并不足够，香港作为一个现代化的国际城市，要有一个具杰出工作能力并获大众支持的特首方是王道。另外，张先生一再要钟博士为过去传媒的报道负责。这显然不是统计问题，但我亦想请教张先生，有几许公众人物包括特首、司长、局长以致阁下又何曾会为传媒的报导负责呢？梁振英 N 年前也说不会选特首，张先生曾几何时亦公开赞扬港大民调中立专业。那张先生又如何对自己的言论负责？梁特首又是否要为自己反口食言负责呢？

事实上，民调是一项以统计学为基础的社会研究专门科学，张先生可能并不是这方面的专才，那么还请留待其他学者们讨论交流。而张先生贵为行会成员，亦请不要重私忘公，免得引起社会大众误会行会打压学术自由，那就相当不妙!

最后，我要向张先生表达敬意，你甘愿接纳与支持一个不足 50 分的特首，只因视特首如己出，把他当作儿子般看待，实有为人母亲的伟大情操！（主席按：果然是温良恭让的学者，未句明明就是「他妈的」伟大！）

港大民调之统计学解读《有涯小扎》

摘要：本文透过统计学分析方法，检视近日舆论对港大民调中特首民望调查的批评及反驳，探讨这些言论背后的统计学理据。本文作者认为，港大民调在抽样方面十分严谨，但在设计问卷和演绎结果方面有值得适榷之处。本文又对港大民研所公布的原始数据进行了进一步分析，指出当中所蕴含的启示，并据此提出建议。

引言

近日有关香港大学民意调查（下称港大民调）的争论甚嚣尘上。港大民调是香港大学民意研究计划（下称港大民研）定期举行的民调，由香港大学政治与公共行政学系的钟庭耀主持。民调内容包括特首、政府、主要官员、议员民望，及其它社会指标等（《香港大学民意研究计划》）。2014 年 2 月 8 日，民主党党员、律师陈庄勤在明报发表《沉默的螺旋》一文，批评港大民调以平均分来表达特首梁振英民望，结果易被极端数值影响，又以 50 分作为合格分数，并不全面。同时这些民调「本身并不单单在反映民意，也同时在以定期公布评分来塑造民意」（2 月 8 日明报陈庄勤《沉默的螺旋》）。3 月 4 日，在北京举行的政协港澳联组会议上，政协常委、恒基地产副主席李家杰点名批评钟庭耀，指其主持的港大民调「总是在关键时候发表对中央政府、特区政府以至整个爱国爱港阵营十分不利的民意调查结果」，藉此「操弄民意」。他又认为钟的民调不够科学，却是香港众多民调机构中最具影响力的一个，必须尽快改变（3 月 5 日 AM730《李家杰批评钟庭耀用民调为反对派造势》）。钟庭耀于同日发表书面声明响应，指出其调查方法经得起学术考验，「总会坚持科学透明的原则，从不迁就对方的政治背景或立场」，认为「如果把言论自由的忧虑，进一步扩大至学术自由的空间，是非常不智的做法。」他又欢迎任何人士讨论民意研究工作，「只要是实事求是，客观公正，便可集思广益」（港大民研《关于政协委员李家杰于政协会议上有关「民意调查」的言论》）。

争论焦点

陈、李二人的批评引起了广泛关注。有论者从政治立场和动机立论（如 3 月 17 日文汇报文平理《「钟氏民调」真的是学术吗？》、3 月 18 日苹果日报李怡《攻民调为扼杀民意》），本文对此无意涉猎。另有论者从统计学角度评论钟的研究方法。行政会议成员张志刚在电台节目称，钟庭耀曾经多次提到 50 分是合格水平，认为他有需要向公众交代（3 月 20 日商业电台《张志刚指钟庭耀多次提及五十分属合格》）。他又认为，在极端评分的影响下，用平均分来评核梁振英表现，犹如瞎子摸象，普通人亦难以理解 50 分是否合格水平。若 50 分属于不合格，港大应清楚说明，并解释何谓支持度评分合格或不合格 （3 月 21 日大公报《张

志刚促钟庭耀交代 民望 50 分是否合格》）。陈庄勤则指出，「在一般人心目中，50 分这及格分具有非常重要的象征意义」，但如果只公布平均分而不公布各评分的人数分布，便是不完整的民调结果公布。以今次民调为例，61.8%受访者给予合格分数，38.2% 给予不合格分数，跟两大民研 / 民调机构定期公布以平均分均多数低于 50 分所显示的民情相去甚远（2 月 8 日明报陈庄勤《沉默的螺旋》、3 月 20 日明报陈庄勤《再谈民调》）。网站「港人讲地」亦提出类似论点，指出整体平均分被 0 分的「极端评分」拉低，令梁振英支持度被低估，认为应取中位数更佳。过往多年的新闻报道都把 50 分演绎为及格分数，港大民研亦未有澄清，令市民累积了「50 分等同合格」的印象。又批评港大以 SPSS 格式发布原始数据，必须装有特定软件才能开启（3 月 14 日港人讲地《解开特首民望「不合格」之谜》、3 月 20 日港人讲地《有关港大民调的几个疑问：覆练乙铮及关焯照两位学者》）。公民党党员、港大法律学院院长陈文敏认为，剔除极端数据是普遍做法，因为更能反映现实（YouTube 视频《公民党港大法律学院院长陈文敏都觉得钟庭耀的民调做法不是专业手法》）。中大亚太研究所研究员郑宏泰称，港大民调的 50 分没有正面意思，不能视为合格，与中大民调讲明 50 分及格并不相同。但 0 分亦是表达出某类民意，从政者应予注意（3 月 20 日明报《特首民望 50 分意义 中大「及格」 港大「一半半」》）。

因应批评，钟庭耀在港大民研网站重贴了 2003 年的两篇文章，解读特首民望调查的设计（《「特首民望新解」、「问责官员如何向民意问责？」》）。文章指出，55 分的支持度大约等如假想投票中的 45％的「得票率」，50 分的支持度则可化成大约 30％的「得票率」，45 分大概会转化为 20％，而 40 分大概会化成 10％至 15％左右。其后，钟又在出席一个论坛时响应，指使用平均分是国际常用标准。而 50 分只是一个中性的分数，没有所谓合格不合格。至于开启 SPSS 格式档案的软件，在大学可以免费下载，他相信任何一个专业研究机构都有相关软件（3 月 15 日商业电台《钟庭耀指国际间最常使用平均分作研究结果》）。前中大经济学教授、现职冠域商业及经济研究中心的关焯照，联同经济学家周文林、统计学家雷照盛等撰文，指出根据问题的措辞，50 分只是代表「一半半」，没有任何暗示这是一个合格的最低门坎。如果把 50 分归入合格，会得出 61.8%的人给了合格分数。但如果把 50 分归入不合格，会得出 66.4%的人给了不合格分数，两者结果相反。解决方法是把一半评 50 分的人归入 0-50 分一组，另一半归入 50-100 分一组，结果是有 52.4%的人给了 0-50 分，反映特首的支持度评分不是太理想。他们同意一旦出现很多人选择极高或极低评分，平均分不是最好的指标，建议同时公布中位数和众数，或剔除极高或低评分部份，计算「截尾均值」。但他们亦认为，极高和极低的评分也是重要的统计资料，不能忽略（3 月 20 日苹果日报关焯照、周文林、雷照盛《民调小学鸡》）。传媒工作者练乙铮则指，港大民调的特首民望评分由 0 至 100，即有 101 个整数，50 分居中，故此应尊重给予 50 分者的中立态度，而非把 50 分理解为支持梁振英。至于 0 分与 100 分，在港大

民调中都有清楚而具体的定义，不应剔除。若真要剔除 0 分，亦应同时剔除 100 分。即使剔除了，平均值仍是低于 50 分（3 月 20 日信报练乙铮《打棍无效：网小子放倒「巨人」张志刚》）。

下表总结了两方面的言论：

	批评	反驳
平均分与极端评分	▪ 整体平均分被极端评分拉低，低估特首支持度。(陈庄勤、港人讲地) ▪ 剔除极端数据是普遍做法，更能反映现实。(陈文敏) ▪ 一旦出现很多人选择极高或极低评分，平均分不是最好的指标。可同时公布中位数和众数，或剔除极高或低评分部份，计算「截尾均值」。(关焯照等) ▪ 类似 0 分或 100 分的的极端评分将会愈来愈多，因此不能单单公布平均分，可以中位数代之。(港人讲地)	▪ 使用平均分是国际常用标准。(钟庭耀) ▪ 0 分亦表达出某类民意，从政者应注意。(郑宏泰) ▪ 极高和极低的评分也是重要的统计资料。(关焯照等) ▪ 0 分与 100 分都有清楚而具体的定义，不应剔除。若真要剔除 0 分，亦应同时剔除 100 分。即使易除了，平均值仍是低于 50 分。(练乙铮)
关于 50 分是否合格分数	▪ 以 50 分为合格分数并不全面。给予合格分数的人数是占总受访人数的 61.8%，给予不合格分数的人数占总受访人数的 38.2%。这样的结果与多年来两大民研 / 民调机构定期公布以平均分均多数低于 50 分所显示的民情相去甚远。(陈庄勤) ▪ 港大民调的 50 分没有正面意思，不能视为合格。(郑宏泰) ▪ 有愈六成人给了 50 分以上的分数。过往新闻报导都把 50 分演绎为合格分数，令市民累积了「50 分等同合格」的印象，港大有必要澄清。(港人讲地)	▪ 50 分只是一个中性的分数，没有所谓合格不合格。(钟庭耀) ▪ 55 分的支持度大约等如假想投票中的 45％的「得票率」，50 分的支持度则可化成大约 30％的「得票率」，45 分大概会转化为 20％，而 40 分大概会化成 10％至 15％左右。(钟庭耀) ▪ 根据问题的措辞，50 分只是代表「一半半」，没有任何暗示这是一个合格的最低门坎。50 分是评分的中间点，如果把 50 分归入合格，会得出 61.8%的人给了合格分数。但如果把 50 分归入不合格，

	▪ 翻查以往报道，发现钟庭耀曾多次提到 50 分是合格水平。普通人难以理解 50 分是否合格水平，认为钟要澄清。(张志刚)	会得出 66.4%的人给了不合格分数，两者结果相反。解决方法是把一半评 50 分的人归入 0-50 分一组，另一半归入 50-100 分一组，结果是有 52.4%的人给了 0-50 分，反映特首的支持度评分不是太理想。(关焯照等) ▪ 特首民望评分由 0 至 100，50 分居中，应尊重给予 50 分者的中立态度，不应擅自将「50 分」定义为「合格」。(练乙铮)
数据格式问题	▪ 港大以 SPSS 格式发布原始数据，必须装有特定软件才能开启。(港人讲地)	▪ 开启 SPSS 格式档案的软件，在大学可以免费下载，相信任何一个专业研究机构都有相关软件。(钟庭耀)

关于民调的统计学基础

民调在外国称为 opinion poll，其要旨是运用统计学方法，找出一个群体对于某个社会议题的意见。统计过程可以分为五大步骤：收集、组织、分析、演绎、发表（《What Is Statistics? – Overview 》)。

做民调的最理想方法是从整个群体（称为「母体 (population)」）中收集数据，即要访问群体内的所有人，如此即能得出全面的统计数据，这种做法称为「人口普查 (population census)」。但现实中往往由于目标群体的人数众多，只能从受访对象之中作随机抽样 (random sampling) 并进行访问，这种做法称为「抽样统计 (sample statistics)」。无论是人口普查或抽样统计，在得到原始数据之后，研究员都会组织并分析原始数据以进行总结。最常见的总结方法是取平均值 (mean) 和标准偏差 (standard deviation)，以展示数据的中央趋势 (central tendency) 和分散程度 (variability)。中央趋势的量度，还可以用中位数 (median) 和众数 (mode)。分散程度的量度还可以用数值范围 (range，即最大数减最细数)、方差 (variance，即标准偏差的平方）、百分位数 (percentile)　等。除了中央趋势和分散程度，有时还要量度数值分布的偏度 (skewness，即非对称性) 和峰度 (kurtosis，即尖峰的尖锐程度)。这些都是尝试用少量的数字，去总结一大堆数据的整体特性。数字之外，有时也会用图表表示数据的特性，最常见的是以直方图 (histogram) 来展现数据的频率分布 (frequency distribution)。从上文可知，数字简洁易用但流于片面，图表表达较麻烦却能给出更多方面的数据，研究员在报告中往往要两者配合使用，才能展现数据的真实特性。

用这些统计结果来描述原始数据的特性，称为描述性统计 (descriptive statistics)。如果是从样本的特性来推论整个母体的特性，则称为推论性统计 (inference statistics) 。中央极限定理 (central limit theorem) 表明，如果样本数足够大，而且抽样足够随机，则样本的平均值会呈正态分布 (normal distribution) 并趋近母体的平均值，而标准偏差则为母体的标准偏差除以样本数的开方。只要符合中央极限定理的条件，便可以从样本的平均值和标准偏差，推测母体的平均值和标准偏差，并推测这些推测的置信区间 (confidence interval) ，以估计可能的误差范围，从而决定推测的可信性。然后，研究员便会就着有关调查的主题，演绎并发表调查结果。

关于上述的统计学理论，可以参考一般的统计学入门书籍（如《OpenIntro Statistics》）。

抽样调查可能出现以下几种误差：
其一、因为样本缺乏代表性而引入误差。抽样必然要忽略母体中部份人士的意见，样本越小，遗漏越多，因此样本必须要有代表性，即其成份跟母体相若，否则从样本的特性来推论整个母体的特性时，便会出现误差 (Wilks, 1940)。例如，有文献指出部份在美国进行的电话调查，只对家用电话号码进行抽样，但现今越来越多人只用手提电话，作者认为有证据显示这些只用手提电话的人有相当不同的政见，因此以家用电话受访的样本不能代表他们 (Mokrzycki, 2010)。

其二、受访者未必愿意表达自己的真实看法。例如问题较敏感，令受访者不想或不敢表达意见。有学者提出沉默的螺旋 (spiral of silence) 的概念，指出如果受访者认为自己的意见属于少数派，便可能不敢发表真实的意见 (Noelle-neumann, 1974)。一项以台湾与美国人为对象的研究指出，接受电话访问时台湾人展现了沉默的螺旋现象，美国人则不然，显示某种文化特质可能会导致这现象出现 (Huang, 2005)。

其三、访问的用语或会影响结果。不同文化、不同背景的人对问题可能有不同的理解 (Groves, 2009)，影响数据的有效性 (validity)。

其四、在总结报告时，无可避免要忽略原始数据中的一些数据。例如平均值的计算方法是将数据总和除以个数，从平均值却不能反过来计算出原始数据。以 {0, 60, 60} 和 {40, 40, 40} 两组数据为例，平均值都是 40。两组数据明显不同，却无法从 40 这个数字得知有甚么不同，因为原始数据的细节被忽略了。如果统计量的选取不宜，便会在演绎出误导的结果。部份舆论针对平均值所提出的质疑，即属这一类。

港大民调使用的方法

港大民研网站详列了特首梁振英评分的相关研究方法（《特首梁振英评分》）。调查基本上每两个月进行一次，以电话访问 18 岁以上操粤语的香港市民。每次样本数为 1000 或以上，抽样方法是从住宅电话簿中首先以随机方法抽取「种籽」号码，在号码上加减 1 或 2，过滤重复号码后再作随机排列，然后提供给访员进行电话访问。如果被抽中的家庭中成员不止一人，就选择下一位即将生日的家庭成员作访问。

调查的结果经过了加权 (weighting) 处理。根据上文所引文献 (Wilks, 1940)，样本的成份要跟母体相若才有代表性。由于事实并不符合这项要求（例如年龄分布不同），因此研究员按 2013 的中期人口统计中的性别与年龄分布，及 2011 年人口普查中的学历分布，对样本进行了加权，其百分比已详列于《被访者基本个人资料》网页。例如，18-29 岁的人口比例，在原始样本中为 15.9%，在加权样本中修正为 18.3%。要留意加权是加在人数上，而不是加在分数上。两者的概念大有不同。例如一个给了 50 分的人，若要将其所占的权重加倍，会变成两个给了 50 分的人，而不是一个给了 100 分的人。有些网站忽略了这一点，错误计算出大于 100 分的评分（如：辅仁网《港大民研特首评分系「被拉高」还是「拉低」？》）。调查所用的问卷有几个版本，关于特首民望的问卷编号为 tp1403013_01（《调查问卷》）。除了询问受访者对特首的支持度之外，问卷还会询问受访者的居住地区、家庭成员人数、是否登记选民、有否在各项选举中投过票、性别、年龄、教育程度、居住情况、婚姻状况、职业收入、阶层（如中产、基层等）、出生地、行业、来港年期等等。

关于特首支持度的问题有两条：

- Q1: 而家想请你用 0 至 100 分评价你对特首梁振英既支持程度，0 分代表绝对唔支持，100 分代表绝对支持，50 分代表一半半，你会俾几多分特首梁振英呢？
- Q2: 假设明天选举特首，而你又有权投票，你会唔会选梁振英做特首？

备受争议的民望评分即来自 Q1 的答案。基于近日公众的关注，港大民研网站公布了最近一次（2014 年 3 月 3 日-6 日）的原始数据，文件格式为 SPSS，内里包含了 Q1 的数据共 1017 条，亦即此次调查的样本数。根据 SPSS 文件内的说明，其数据结构如下：

- 第一列：1-1017 的编号；
- 第二列：受访者所给的 Q1 的分数；其中 3 条记录是 191，代表「不认识梁振英」。16 条记录是 8888，代表「不知道」或「不肯讲」。余下 998 条为 0-100 间的整数，即为受访者给予梁振英的评分。

- 第三列：性别；其中 1 代表男，2 代表女。
- 第四列：年龄组别；其中 1 代表 18-29，2 代表 30-39，3 代表 40-49，4 代表 50-59，5 代表 60-69，6 代表 70 或以上。另有 4 笔记录是-99，代表拒答。
- 第五列：一个代表权重的数字；例如第一笔记录的人的权重是 0.85422675557，表示他在经加权处理的样本中，只代表 0.85422675557 个人。

就着 Q1 的答案，港大民研原先发表的报告中只报告了以下数点（《港大民研发放特首及问责司局长民望数字》）：

1. 特首梁振英的最新支持度评分为 47.5 分，跟两星期前变化不大。
2. 样本数是 1017。
3. 回应率是 65.9%。
4. 误差率是 +/-1.5，即 3% （以 95%置信水平计算）

注：报告亦提及，根据民研计划的标准，梁振英属于「表现失败」，其定义为反对率超过 50%。但反对率来自 Q2 的答案，不在本文讨论范围内。有论者认为「表现失败」是因为梁的平均分在 50 分以下，从而引发关于定义合格分数的批评。按照调查中所用的「民望级别总表」中的定义，这项批评并不符合事实。

分析及评论

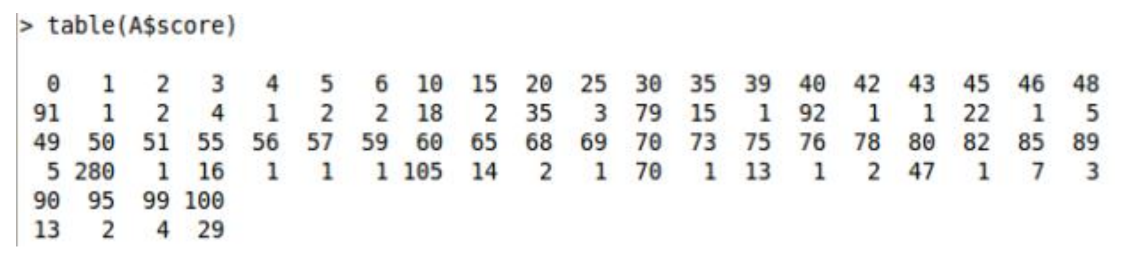
参照前述抽样调查可能出现的几种误差，比较港大民研网站所列的研究方法、数据和分析，我们可以评价港大民调在特首民望评分上面的合理与否。

港大民调以电话进行随机访问，对种籽电话号码进行加减处理，并以生日日期选取家庭成员作访问。最终成功访问的样本数达 1000 以上，响应率 65.9％，又对数据进行加权处理，应能很大程度上确保了样本的代表性。以家用电话号码来抽样，可能会出现美国研究中描述的偏颇情况。但目前没有证据显示，忽略手提电话的用户会对关于特首民望的调查造成偏颇的结果，因此不能以此作为对港大民调的指控。

文献指出人们可能会因为自己的意见属于少数派而不敢发表真实的意见，即「沉默的螺旋」现象。但是次电话访问以匿名进行， 应能减低人们的担忧。而且即使「沉默的螺旋」存在，除非人们认为大多数人都很极端，否则「沉默的螺旋」亦只会令人们倾向选取中间的答案，不会反过来导至「极端答案」的出现。

访问用语方面，问卷的说明是 0 分代表绝对不支持，100 分代表绝对支持，50 分代表一半半。如果受访者要从这三个分数中选择，大部分都会选中间的 50 分。如果要给其它分数，受访者就要思考其它的数字。图一显示各分数的出现频率，图二将这频率以图象方式表达。从这些数据可知，受访者倾向给出简单的数，其中 0 字尾的数字最多（如 0,10,20,30,…），5 字尾的数字较少，其它数字最多只有

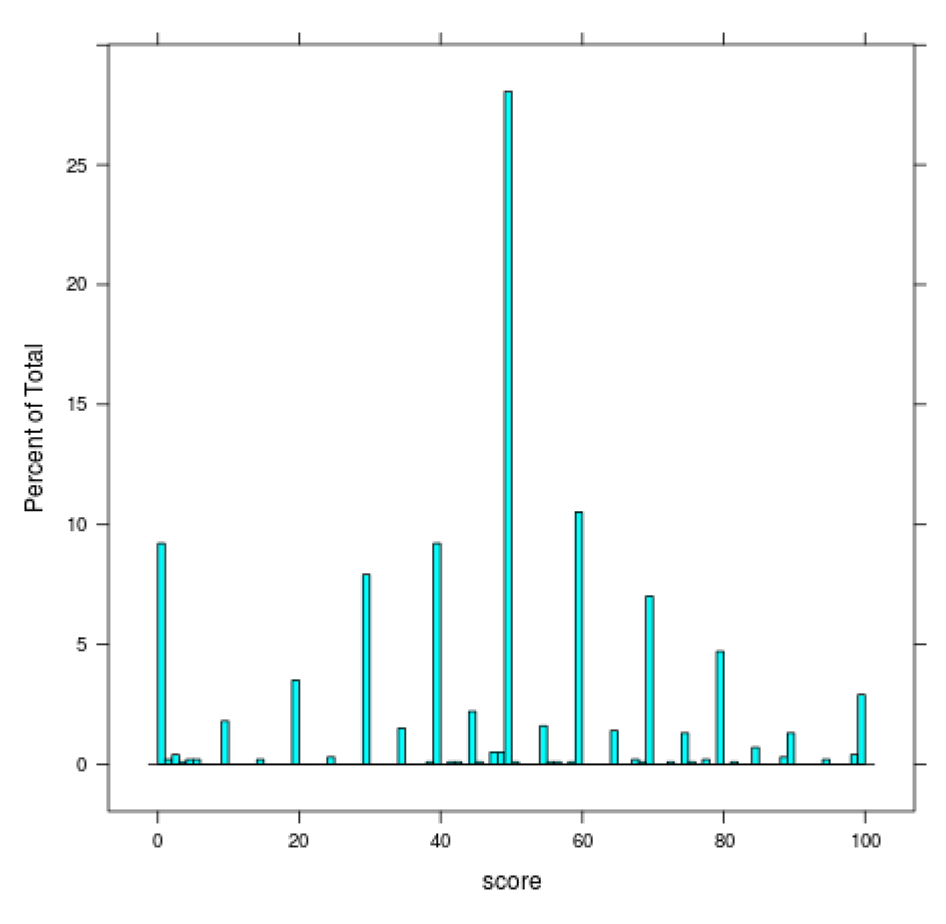
几个人选择。另外，选 50 分的人非常多，共 280 人，选 0 分的有 91 人，选 100 分的也有 29 人。这三个分数的出现频率比旁边的分数多出很多。理论上，1 分甚或 10 分的相差应该算是轻微的变化，但对受访者来说，这 0,50,100 三个分数都具有独特意义。1 分跟 2 分之间可能没有差别，0 分与 1 分之间的差别却是巨大的，是质变而非量变。同理，100 分与 99 分之间，49、50、51 分之间的差别亦然。民调要求受访者给出 0-100 之间的分数，并以此计算平均值，是假定了这个分数跟受访者心目中对特首的支持度之间有一连续变化的线性关系。事实上，问题的问法赋予了三个分数特别的意思，客观上扭曲了分数分布。这效应在 50 分这一临界点尤为重要，下面再详述。



```
> table(A$score)

  0   1   2   3   4   5   6  10  15  20  25  30  35  39  40  42  43  45  46  48
 91   1   2   4   1   2   2  18   2  35   3  79  15   1  92   1   1  22   1   5
 49  50  51  55  56  57  59  60  65  68  69  70  73  75  76  78  80  82  85  89
  5 280   1  16   1   1   1 105  14   2   1  70   1  13   1   2  47   1   7   3
 90  95  99 100
 13   2   4  29
```

图一：各分数的频率分布

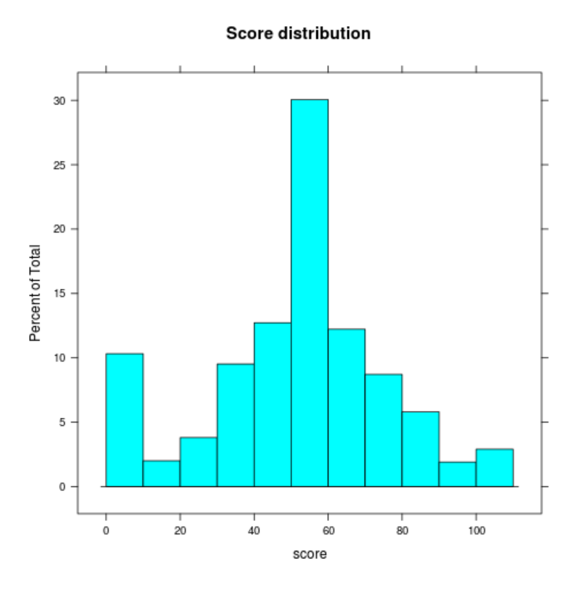


图二：分数的频率分布图（以 1 分为一格）

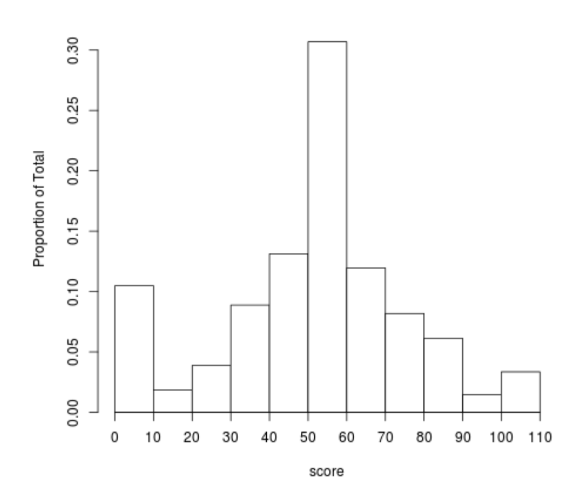
原报告以报导平均分为主，新闻媒体主要亦以这个数字作为讨论的根据。如前所言，平均分只是总结统计数据的其中一种方式，不同的统计量会给出不同方面的信息。平均分是最常用的方式，其好处是计算涉及所有的数据，坏处是易受极端数字影响。如果数据中出现极端的数字，一般做法是以中位数取代。中位数是指将数据顺序排列之后排在中间的数。例如，数集 {0,0,0,0,100} 的平均值是 20，中位数是 0。平均值因受 100 影响，其数值不能很好地反映数集的中央趋势。反之，中位数只取决于数字的排列，在这情况下就较能反映中央趋势，这就是为甚么入息通常都是以中位数而非平均值来计算中央趋势。至于众数，则是频率最高的数，在这例子也是 0。也有一些情况是三个数字都不能很好地反映中央趋势。例如，数集 {0,0,0,100,100,100} 的平均值是 50，中位数是 50（中间两个数的平均），众数是 0 和 100 （因频率相同），三个数字都难以代表数集的总体特性，因为数集本身就是分化成两边的。一般来说，只有当分布接近钟形分布时，这三个统计量才能较好地反映现实。

从原始数据可知，是次民调的分数分布并不依从钟形分布，单纯从数字很难对统计结果作出全面的认识，因此以下改由图表进行分析。

图三是以每 10 分为一组的频率分布，是未经加权处理的结果，分组方法为 0-<10、10-<20、20-<30、30-<40、40-<50、50-<60、60-<70、70-<80、80-<90、90-<100，100-<110。留意最后一个分组实际上只有 100 分的分数。一般做法是把 100 分归入前一组，变成 90-100。但因在这组数据中，100 分出现了峰值，所以做了这个特别处理，以免影响了前一组的结果。加权处理则按各权重调整每一组的频率，分组方法相同，结果如图四所示。



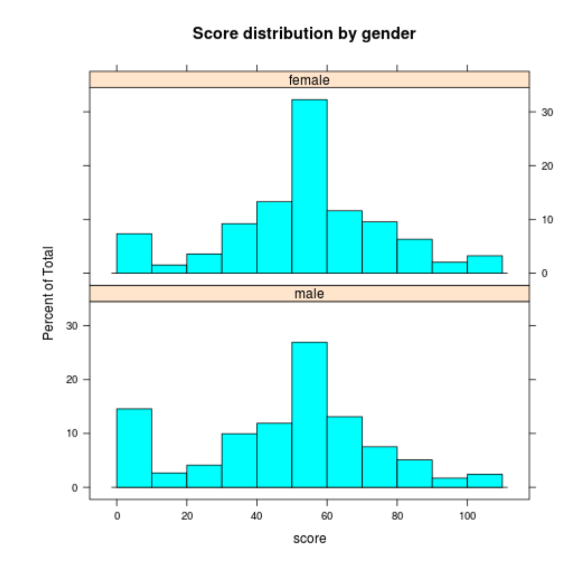
图三：未经加权处理的频率分布



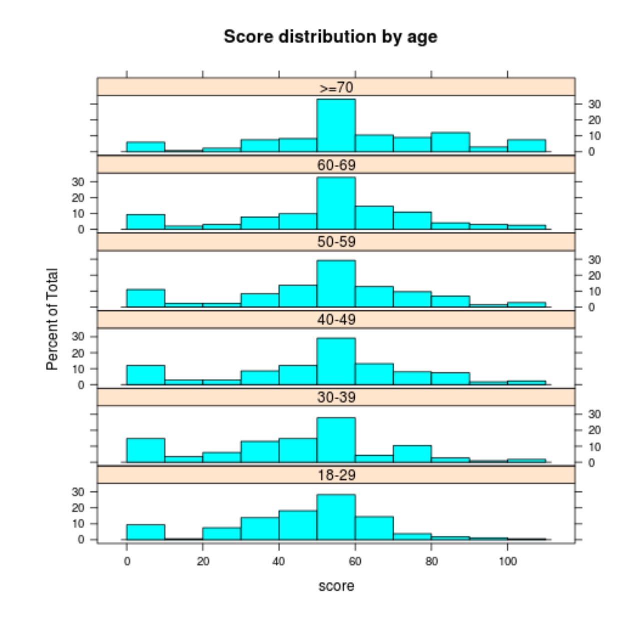
图四：经过加权处理的频率分布

两幅图只有些微差别。由于本文的分析以看图表为主，不涉及计算合格不合格的问题，为了方便说明，以下将采用未经加权处理的频率分布。

跟图二的结果一样，图三清楚展现了 0 分、50 分和 100 分的特殊性。除了总体的分布外，港大公布的原始数据还包括年龄和性别的资料，因此我们也可以按性别和年龄分别画出各组别的分布，如下面两幅图所示。



图五：以性别分组的分数分布



图六：以年龄分组的分数分布

先看 0 分的情况。无论是按性别还是年龄分组，都可以看到 0-<10 分处出现尖峰。从原始数据或图二都可以看出，在这个组别里绝大部分都是直接给了 0 分。进一步说，男性受访者给 0 分的人较女性多，有接近 15%。而 30-39 岁的组别给 0 分的人较其它组别多，亦是接近 15%。从 40 岁开始，年纪越大的组别，越少人给 0 分。即使忽略了这些给 0 分的情况，也可以看出 18-29 岁及 30-39 岁的市民，评分少于 50 分的较评分多于 50 分的为多。而随着年纪增加，排除 0 分之后两边趋向平衡。到了 60-69 岁及 70 岁或以上的组别，则有向右边发展之势。因此，如果以给 0 分的作为对特首极度不满的标示，则可以看出最不满特首的是介乎 30-39 岁的市民。从 40 岁的组别开始，年纪越大的市民对特首的支持度越高。18-29 岁是刚刚毕业出来工作的年纪，30-39 岁是成家立业的年纪。这两个年龄层的不满，或许反映了政府在经济、就业等政策上的不足，也有可能是这个年龄层的人较关心政治，尤其是在民主发展上产生不满。真正原因必须经进一步研究确定，本文只能从数据上指出这一现象，没有足够的资料作出解释。

再看 50 分和 100 分的尖峰。明显的 100 分尖峰只出现在 70 岁或以上的组别。事实上，70 岁或以上的组别，50 分尖峰两边的分布很均匀，而 50 分尖峰比其它组别都突出。图二的分布也显示，50 分尖峰的人数，远远超出了钟形分布应有的数量。透过比较旁边两组的高度，大约也是多了 15%。如前所述，问题的设计很容易令人选择 50 分。这些人要么真是觉得自己对特首的支持度是一半半，也有可能只是觉得难以下决定，或者根本没有打算认真思考这个问题，只好给一个中间的分数。如果这班人经过了详细思考，就可能会给出较高或较低的分数。鉴于这班人的人数不少，他们的决定会对整体分布产生关键影响。无奈问卷的设计无法把这批人分辨出来，因此我们不知道这班人的真正取态。

总结及建议

本文透过统计学分析方法，尝试检视近日舆论对港大民调的批评及反驳，探讨这些言论背后的统计学理据。本文作者认为，港大民调在抽样方面十分严谨，但在设计问卷和演绎结果方面有值得适榷之处。

其中，无论以平均分、中位数还是众数来进行统计，都不能全面地反映调查结果。应该同时公布频率分布，甚至是各年龄组别的频率分布，才能从中提出改善施政的建议。在分析极端分数的时候，我们可以把这些分数分开来考虑，以反映其他人的意见，但极端分数还是有它的重要价值。至于给予 50 分的人数众多，本文认为是来源于问卷设计出现了问题，致使难以得知这些人的真正取态，降低了调查的价值。

关于合格分数的问题，由于原问卷设计中，50 分只是一半半的意思。以 50 分为合格分数可能符合一些人的直觉，但本文认为没有压倒性的理由以此定义为合格分数。合格是最低要求的指标，但这个最低要求设在何处则是没有一定准则。即使在学校的考试制度里，合格分数也并非每间学校相同，只能说通常在 40-60 分之间。本文同意钟氏的说法，50 分只是一个中性的分数，没有必要跟合格不合格挂钩。传媒亦不应再以此作为报导的焦点。

此外，从按年龄组别画出的分数分布可以看出，民调的数据确能反映一些重要的社会现象 。虽然大多数人中间落墨，所谓的极端分数只占少数，但亦有一成之众，而且集中在 30-39 岁的组别。在一个社会里，沉默的大多数和激进的极少数同样重要。前者是社会稳定的要素，后者是变革的动力，缺一不可。为甚么某些组别的人给了最差的评分，他们最关注的是甚么，这方面的跟进工作，不但能够响应这组人的关注，亦有可能带动社会的整体进步，从政者责无旁贷。

最后，本文作者很感谢港大民研公开最近一次民调的原始数据，让社会大众可以进行更深入的分析。然而 SPSS 只是学术界常用的统计软件，但如果数据的使用对象是传媒或一般大众，通常的做法是一并提供 CSV 和 Excel 版本，有时也会提供 XML 版本（参看：美国政府的《Data.gov》、香港政府的《资料一线通》）。现时在 MS Excel 上开启 SPSS 格式档案必须另外安装插件，本文作者亦是使用了 PSPP （《PSPP – GNU Project – Free Software Foundation》）或在 R （《The R Project for Statistical Computing》）安装某些特定的程序包才能开启。若能以比较普及的格式提供数据，将有助信息的透明和公开。

其他参考数据

华生：〔中国〕城乡差距的统计误导和真实挑战
媒体称统计造假误导决策 病根在于数字出官
人民日报批评地方 GDP 报花账：误导宏观决策
但愿失真的统计数据不会误导个税决策
林美芬：内地数据「灌水」 误导中央害经济

骗人的诚实数字：谈谈《欧盟动物园报告 2011》于对比圈养及野生海豚死亡率数据时所做的误导

潘震泽：民调可靠吗？

萧亮思：为甚么「小学鸡统计」应纳入通识？